

# How not to prioritize

A high-level reprimand to US astronomers highlights the need for the objectives of 'big science' to be openly debated.

On 8 January, NASA's administrator Mike Griffin upbraided the annual meeting of the American Astronomical Society in Austin, Texas, for a lack of team spirit. Not only were some of the scientists less than enthusiastic about the human-exploration goals that have been this administration's top priority in space, they were also messing up the agency's astrophysics programme with special pleading to Congress. Such interventions, Griffin warned, would thwart the community's stated goals and blight its future projects.

The meatiest bone of contention was the Space Interferometry Mission (SIM). SIM offers a way of discovering planets by observing slight jitters in the position of their parent stars. NASA has spent nearly \$600 million on it already. On the basis that finishing it might easily cost a further \$1.85 billion (see page 228), the agency had planned to assign it \$22 million in this year's budget, keeping it far from any prospects of flight. Congress gave it three times that much, apparently wishing to see it move into full development. Such development would, warned Griffin, leave NASA no room for any other astrophysics missions of any size, and force delays or cancellations on those already in development.

Griffin sought to portray the boost for SIM as a fratricidal move to circumvent the settled result of the astronomers' 'decadal survey' process. Under the auspices of the National Academy of Sciences, the community gives its funding agencies, and the lawmakers who provide their budgets, regular surveys of its priorities. This process has been much praised, but is marred by shortcomings. In particular, the most recent survey was undercut by a self-deluding ineptitude on matters of cost. The James Webb Space Telescope was just one of the seven 'major initiatives' prioritized by that survey, yet by the time of its launch it will on its own have cost far more than the total that the survey envisaged for all of them. As that most recent wishlist also included spending on as-yet unfinished projects from the previous survey (of which an early form of SIM was one), it is not clear what help it offers decision-makers today.

Another problem is that there are legitimate interests in the future of American space research that such surveys may not capture. SIM is a project based at the Jet Propulsion Laboratory (JPL) in Pasadena, California. The advent of full-cost accounting at NASA, which means that money follows specific projects to a greater extent than ever before, has heightened the importance of flagship projects for the institutions that host them. A healthy budget for SIM could serve to maintain a pool of talent at JPL that might otherwise be eroded; if you want a reason for the lobbying, this is a pretty good starting point. The beauty of such a power-house is no doubt in the eye of the beholder, but everyone should recognize that the benefits of a healthy JPL are felt beyond the precincts of Pasadena.

The lesson for the astronomical community is that the decadal survey should provide a range of more and less capable missions, thus making it easier for policy-makers simultaneously to satisfy the community's goals and the constraints of the public purse. It should also agree that the imprimatur of priority bestowed by a decadal survey has a use-by date — after a certain time, perhaps as little as five years, it is reasonable to ask whether a given mission is still the best way to achieve its stated goal. Anyone setting priorities needs to scrutinize SIM in this spirit.

Meanwhile, NASA's administrator needs to accept that Congress has a legitimate role in setting goals for his agency. He should also consider that portraying the astrophysics budget as a zero-sum game is a tactic that could backfire: if astronomers thus threatened successfully lobby for a significant transfer of funds from human spaceflight to science, his position will be weakened. And Congress should, when exercising its powers, open up a public debate on all the issues involved — which may often go beyond the merits of a single mission. ■

## Deserting the hungry?

Monsanto and Syngenta are wrong to withdraw from an international assessment on agriculture.

"This is a most reluctant decision." These are the words of a spokesman for the agriculture-industry body CropLife International speaking to *Nature* this week. The decision in question is that by two CropLife member corporations, Monsanto and Syngenta, to pull out of the International Assessment of Agricultural Science and Technology. This is an ambitious, 4-year, US\$10-million project that aims to do for hunger and poverty what

the Intergovernmental Panel on Climate Change has done for another global challenge.

The scale of the ambition is clear both in the project's promised outcome, as well as in its internal workings. When published later this year, its reports promise to map how science, technology and accumulated good-farming practice can be used to reduce hunger and improve quality of life for rural people in developing countries (drafts can be accessed from [www.agassessment.org](http://www.agassessment.org)). At the same time, the writing and review teams (some 4,000 experts in all) comprise a grand coalition including scientists, government officials, representatives from seven UN agencies, farmers' groups, a rainbow of non-governmental organizations (NGOs) and industry, including chemicals manufacturer BASF and agri-biotech giants Monsanto and Syngenta.

But these last two, part of the assessment from the beginning, have now decided to quit. No public statements have been offered, but the spokesman for CropLife told *Nature* that the decision was prompted by the inability of its members to get industry perspectives reflected in the draft reports. One of these perspectives is the view that biotechnology is key to reducing poverty and hunger, and it is based in part on high (and rising) levels of demand for biotech crops from farmers across the developing world.

Insiders agree that the current draft is decidedly lukewarm about the technology's potential in developing-world agriculture. The summary report, for example, devotes more space to biotechnology's risks than to its benefits. The report says that evidence that biotech crops produce high yields is not conclusive. And it claims that if policy-makers give more prominence to biotechnology, this could consolidate the biotech industry's dominance of agricultural R&D in developing countries. This would affect graduate education and training, and provide fewer opportunities for scientists to train in other agricultural sciences.

CropLife says that it does not take a "dogmatic" position and remains open to rejoining the assessment if the other team members are willing to be more even-handed. But the views outlined in the draft chapter on biotechnology, although undoubtedly over-cautious and unbalanced, nonetheless do not represent the rantings of a fringe minority. The idea that biotechnology cannot by itself reduce hunger and poverty is mainstream opinion among agricultural scientists and policy-makers. For example, biotechnology expansion was not among the seven main recommendations in *Halving Hunger: It Can Be Done*,

a report commissioned by former UN secretary-general Kofi Annan. The writing team for this report included Kenya's Florence Wambugu, perhaps the strongest proponent for biotechnology in Africa.

The assessment's secretariat and chairs, too, need to ask themselves some searching questions. For starters: how come these founding members of the assessment got to the point of walking out? This is not the first time an initiative has sought to find common ground between

NGOs and industry on a major issue involving science and public policy. There are many lessons that can be learned by talking to, for example, the organizers of the Mining, Minerals and Sustainable Development project, or the World Commission on Dams, both of which produced consensus reports that have had far-reaching impacts.

Whatever happens next, the status quo is not an option. A meeting to agree the final text is expected to take place in April. Monsanto and Syngenta must get back to the table before then. If they maintain their current position, it will be a blow to the credibility of an important scientific assessment. In addition, public confidence in the biotech industry and in its ability to engage with its critics will have been undermined.

Perhaps most important of all, believing as they do that biotechnology is an essential response to hunger, the two companies will be letting down those that they most want to help. ■

**"If Monsanto and Syngenta maintain their current position, it will be a blow to the credibility of an important scientific assessment."**

## Philanthropy needed...

... to save a historic home of scientific stimulation.

**T**he Ciba Foundation's biomedical symposia, which began in 1950, were scientifically influential but were special in other ways too. Thirty or so scientists would gather at 41 Portland Place, part of a beautiful eighteenth-century mews in central London, to spend three intense days discussing a cutting-edge theme, eating formally together in the antique-strewn dining room, and sleeping in overheated bedrooms that cannot be locked from the outside — gentlefolk, after all, don't steal.

Non-British delegates would be bemused by the crazy plumbing and disconcerted by how loudly the undulating floorboards creaked. But all were charmed. Many Nobel laureates have acknowledged the intellectual stimulation of the meetings. Ulf von Euler, for example, said that his ideas of how neurotransmitters are stored and released were stimulated by the foundation's meeting on adrenergic mechanisms in 1960. In contrast, Arvid Carlsson was devastated when the same colleagues rejected his notion — which won him the 2000 Nobel prize — that dopamine was a neurotransmitter in the brain. A reputation could stand or fall on the consensus of a Ciba Foundation meeting.

Time moves on. The Swiss pharmaceutical company Ciba-Geigy was merged into Novartis in the mid-1990s, and the foundation was duly renamed. In 2002, when the foundation's sponsor moved its R&D centre from Basel to Boston, it decided that this old-fashioned, eccentric elegance did not fit its style of conference support (see page 233).

That was a blinkered decision, given the strong links that the pharmaceutical industry needs with the academic community. True, corporate sponsorship has got tougher, with shareholders demanding much greater, and more immediate, accountability. Nevertheless, Novartis should have negotiated more sympathetically with the foundation to explore new approaches. It is easier to destroy an organization with a strong reputation and institutional knowledge than to build one up from scratch.

But the foundation must shoulder blame too. Although it maintained the quality of its meetings, it made no noticeable acknowledgement that the conference game has changed. Its paper-based approach to publication seemed increasingly quaint, for example. Furthermore, the foundation's trustees and directors should have put up a stronger and more public fight for its life. Their decision to work discreetly on the basis of contacts rather than embark on an 'undignified' campaign to find a new sponsor was almost certainly wrong. Be that as it may, the Novartis Foundation is set to be dissolved at the end of next month, having had 15 months to wind things up.

But it is not necessarily curtains for 41 Portland Place. The likely new tenant, the Academy of Medical Sciences, may still be persuaded to continue the international meetings if the right sponsor were to emerge. Such a rescuer might be institutional, as was happily found last year by the similarly small and intense Berlin-based Dahlem Conferences, saved by the newly created Frankfurt Institute for Advanced Studies. Or there may be an enlightened wealthy individual willing to foster top-level scientific brainstorming and debate.

Any offers? ■

# RESEARCH HIGHLIGHTS

## ASTRONOMY

### Old bulk

*Astrophys. J.* **672**, 146–152 (2008)

Massive galaxies are common in the younger reaches of the Universe, the results of a series of recent mergers of smaller galaxies. But some massive galaxies are very old, and probably formed through the rapid gravitational collapse of enormous clouds of gas, Alan Stockton of the University of Hawaii and his colleagues conclude.

Stockton and his team analysed data from the Hubble telescope. They determined the structure of two distant, massive galaxies that seem to have formed early in the history of the Universe.

The disk-like shapes they report signal the collapse of large masses of gas, and are unlikely to have survived large galactic mergers. This implies that some massive galaxies are not merely amalgamations of smaller ones.

## CANCER BIOLOGY

### Arrested development

*Cancer Cell* **13**, 69–80 (2008)

A molecular switch silences a neural development gene in the most common type of brain cancer, according to Howard Fine and his co-workers at the National Institutes of Health in Bethesda, Maryland.

The researchers isolated tumour-initiating stem-like cells from adults with glioblastoma. Some of the tumour cells behaved similarly to stem cells that are destined to become neurons in very young mouse embryos. Specifically, they did not produce a key protein called BMP receptor 1B, which enables cells to pick up external molecular prompts instructing them to keep developing.

Blocking the expression of BMP receptor 1B creates cells that are able to divide but not

## Southern melt

*Nature Geosci.* doi:10.1038/ngeo102 (2008)

In 2006, Antarctica lost three-quarters more ice than it did a decade earlier, researchers have found.

A comprehensive study of the continent's total ice balance concluded that, during the past 10 years, accelerating loss from melting and sliding glaciers (shown in red) greatly exceeded gains from snowfall, which increased in some regions (blue). Both effects are

associated with global warming.

Eric Rignot at the University of California, Irvine and an international team used radar interferometry data to work out glacial flow rates in 1996, 2000 and 2006 along 85% of Antarctica's coastline. The authors also modelled these glaciers' varying thicknesses, allowing them to calculate

the mass of ice lost to the ocean over

time. They then subtracted this figure from the patchy accumulation of the snowpack.

Antarctica's 2006 net ice loss of almost 200 billion tonnes is comparable to Greenland's annual loss, which has been the focus of much discussion about sea-level rises.

to differentiate. Fine and his team report that cells from many of the tumours they analysed shared this molecular glitch.

## BOTANY

### Flower power

*Am. Nat.* **171**, 1–9 (2008)

Interactions with other plant species may influence the arrangement of flowers' structures, researchers have found.

Although it is well established that pollinators can shape flower evolution, the effect of neighbouring plants has remained unclear. Robin Smith and Mark Rausher of Duke University in Durham, North Carolina, investigated the relationship between two species of morning glory, *Ipomoea hederacea* (pictured left) and *I. purpurea*. Pollination of *I. hederacea* flowers with *I. purpurea* pollen yields hollow seeds that do not produce viable progeny, reducing the plant's overall fitness.

The researchers found that *I. hederacea* plants grown in contact with *I. purpurea* flowers showed considerable variation in the arrangement of the flowers' reproductive

structures. Plants with flowers in which the anthers and stigma were clustered closer together produced more seeds than those that held their reproductive parts further apart. The authors observe that this arrangement favours self-fertilization, lessening a flower's probability of being pollinated by *I. purpurea* pollen.

## SELF-ASSEMBLY

### Snakeskin nanobelts

*Nano Lett.* doi: 10.1021/nl0722830 (2007)

Nanoscale algorithmic self-assembly, in which the molecular components of structures are programmed to stick together according to simple rules, could eventually lead to new forms of molecular computing.

Satoshi Murata of the Tokyo Institute of Technology and his colleagues have created DNA 'tiles' that spontaneously clip together in solution, producing ribbons with constant widths of about 100 nanometres.

The self-assembly is defined by sequences of single-stranded DNA on the tiles' edges. When complementary DNA strands match,



R. SMITH

E. RIGNOT ET AL.

the tiles stick together, and the widths of Murata's ribbons are kept in check by special 'boundary' tiles.

The researchers have also programmed the tile-matching rules so that they embody computational cellular-automaton models. The arrangements that result resemble snakeskin belts under the microscope.

## IMAGING TECHNIQUES

### Heliomicroscopy

*J. Microsc.* **229**, 1–9 (2008)

Electrons are commonly used to image materials at high resolution, but their negative charge and high energies can damage fragile samples. To get around this, a group of physicists used helium atoms instead, and successfully photographed a hexagonal copper mesh.

Bodil Holst at Graz University of Technology in Austria and her colleagues propelled helium through a nozzle and used a device known as a Fresnel zone plate to focus the beam onto the copper. This created an image with 2-micrometre resolution.

The experiment demonstrates that helium atoms can generate a picture even when fired at a sample much more slowly than would be required for electrons to produce an image. Holst says the technique might one day be used to image proteins and weak polymers.

## MOLECULAR BIOLOGY

### How to host HIV

*Science* doi: 10.1126/science.1152725 (2008)

To be able to infect human cells, HIV requires more than 250 host proteins, say researchers at Harvard Medical School in Boston, Massachusetts. Only 13% of these proteins have previously been implicated in HIV infection, and the collection could yield potential drug targets for anti-HIV therapies.

Stephen Elledge and his colleagues turned down the expression of more than 21,000 genes in human cell cultures. Each gene was silenced individually in a separate cell line, and all the lines were then tested for their ability to support HIV infection.

The proteins not previously known to have a role in HIV infectivity include some that transport vesicles between organelles, and components of a protein complex called Mediator, which regulates gene expression.

## EVOLUTIONARY BIOLOGY

### A twist in the tale

*Biol. Lett.* doi:10.1098/rsbl.2007.0602 (2008)

A snail with a shell that coils in four directions has been discovered in Malaysia. Reuben Clements of the conservation group WWF-Malaysia in Selangor and his team have described 38 examples of the gastropod — all with curves in similar positions — found in soil from a single limestone site. The creature came as a surprise because the majority of land snails' shells twist around one or two axes. Most species in the genus *Opisthostoma*, in which the new specimens fall, have three coiling axes.

*Opisthostoma vermiculum*, or 'little worm', as the authors have named the curvy creature (pictured below left), is the first of two species with bizarrely arranged coils that the team found.



R. CLEMENTS

## GENETICS

### Lethal matings

*Science* doi:10.1126/science.1151107 (2008)

When two strains of *Caenorhabditis elegans* mate, one-quarter of their grandchildren die during early development because of a weird genetic incompatibility that is maintained by natural selection.

Hannah Seidel, Matthew Rockman and Leonid Kruglyak at Princeton University in New Jersey,

who discovered the incompatibility, crossed worms of the 'Bristol' strain with individuals from the 'Hawaiian' strain, then allowed the offspring to self-fertilize. Those embryos that lacked a gene called *zeel-1* — a deletion characteristic of the Hawaiian strain that is passed on in mendelian ratios — were sensitive to the product of another gene that is carried in sperm. A version of the latter gene from the Bristol strain arrested the development of such embryos.

Because Hawaiian and Bristol worms live together all over the world, the team propose that the incompatibility is not an example of incipient speciation. The genes involved probably confer some unknown benefit to counteract the reproductive cost, they add.

## JOURNAL CLUB

Vivian G. Cheung  
Howard Hughes Medical  
Institute, University of  
Pennsylvania, USA

### A geneticist reflects on DNA sequence variants that influence gene expression and disease risk.

Most people are familiar with the Human Genome Project and the HapMap, which catalogued the millions of DNA-sequence differences among humans. But which of these differences influence our risk of developing diseases remains unclear. This is particularly true for disorders such as heart disease that involve not only many genes but also the interactions among them. In addition, the effects of variations in DNA sequence are often subtle, such as altered levels of gene expression. Identifying those DNA sequences that determine levels of expression across individuals could have great medical potential.

One paper that illustrates this point looks at the two major contractile proteins of the human heart, the  $\alpha$ - and  $\beta$ -forms of the myosin heavy chain (E. van Rooij *et al.* *Science* **316**, 575–579; 2007). Here, Eric Olson and his team at the University of Texas in Dallas identify a microRNA, called miR-208, that regulates how much of the  $\beta$ -form heart cells produce.

A healthy heart requires a particular ratio of  $\alpha$ - and  $\beta$ -heavy chains for its cells to function normally. When stressed, heart cells tend to make too much of the  $\beta$ -form, causing the organ to enlarge, replete with fibrous connective tissue, and less able to contract. This often happens in people with heart disease.

In finding miR-208, the researchers have determined a key component in the molecular basis of heart failure. The next step might be to look for sequence variants of miR-208 and of other gene-expression regulators that could explain why some people are more susceptible to heart disease than others. In this way, whole biological networks could be pieced together and common medical problems more fully understood.

Discuss this paper at <http://blogs.nature.com/nature/journalclub>

## NEWS

# Funding edict for mission has NASA over a barrel

Astronomers in the United States are up in arms after Congress told NASA that it must spend \$60 million next year building a controversial planet-hunting telescope. NASA says the money, nearly three times the \$22 million it had earmarked for the project, will have to be siphoned from the budgets of other missions.

"I hope this is what you want," an inflamed Mike Griffin told the community, "because it appears likely to be what you will get." Griffin, the NASA administrator, was speaking on 8 January at a meeting of the American Astronomical Society in Austin, Texas. With the agency forced to beef up its financial commitment to the Space Interferometry Mission (SIM), there may be a two-year delay to Hubble's successor, the James Webb Space Telescope (JWST). And other future flagship missions to study dark energy, gravity waves and X-ray astronomy might be cancelled altogether, warns Jon Morse, director of the agency's astrophysics division.

The rise, fall, and now rise again of SIM reflects the ongoing debate over the best way to search for an Earth-like planet beyond our Solar System. Earlier versions of SIM were nominated in 1990 and 2000 in community assessments of astronomy priorities, but some wonder how the mission should be re-evaluated as the next decade's priorities are set. And the unusual involvement of Congress only complicates matters. "Congress does not dream up such direction on its own," Griffin points out. "Clearly, external advocacy for SIM has been successful."

Such advocacy is not a secret; nearly all major research institutions have a presence on Capitol Hill. SIM is managed by the Jet Propulsion Laboratory (JPL) in Pasadena, California, which, as a NASA research centre, is forbidden from directly lobbying Congress. But the lab's operator, the California Institute of Technology, also in Pasadena, can. It has previously employed Washington-based Lewis-Burke Associates to lobby for it.

Certainly, someone was able to bend the ear of Adam Schiff, a Democrat who represents Pasadena in the House of Representatives. Schiff is on the subcommittee responsible for funding NASA, and he was instrumental in pushing through the language specifying \$60 million for SIM, saying the project is too

important scientifically for NASA to kill it. "Congress is not willing to take a back seat on this," Schiff says.

SIM's goal is to find planets by using interferometry, which analyses combined waves of light from multiple telescopes. With a 9-metre separation between two telescopes, SIM would make measurements so precisely that it could detect an Earth-mass planet orbiting a Sun-like star at the Earth-Sun distance, as long as the star was within about a dozen light years of Earth.

SIM is by now the result of several generations of discussions about how best to fly an interferometer in space. NASA has been testing the necessary technologies since 1996, but

in the meantime several other planet-hunting missions have moved forward — including France's COROT, which looks for planets passing in front of

their stars; NASA's Kepler mission, to launch in February 2009, using the same method; and Europe's interferometry mission Gaia, which aims to launch in 2012. NASA also has a mission called Terrestrial Planet Finder on the back burner, which Kepler and SIM would theoretically pave the way for. An upcoming NASA task-force report on the best way to look for extrasolar planets will recommend a mission such as SIM as a priority.

But the costs for SIM have escalated. In 2001, JPL thought SIM could be built and launched for \$600 million. But by the end of 2007, nearly that much had already been spent without any building having started. The current best estimate for the remaining mission cost is \$1.85 billion. Realizing that launching the JWST and servicing Hubble could consume the astrophysics budget by themselves, NASA

**"Congress is not willing to take a back seat on this."**



NASA's Mike Griffin will have to reprioritize.

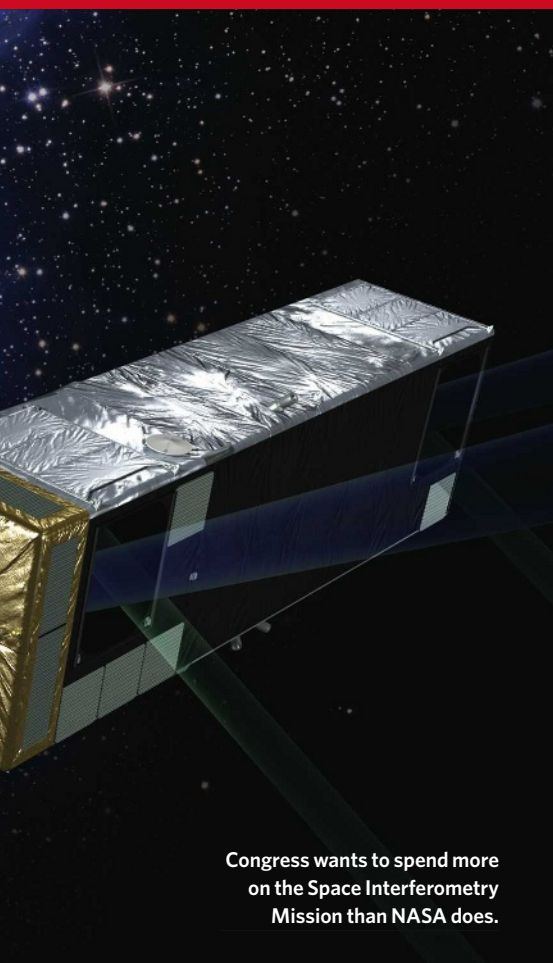


in 2007 began the process of putting SIM into hibernation. Last year, just \$31 million was spent on the programme, compared with \$117 million the year before.

SIM's chances of salvation may lie in re-inventing itself at a much smaller size. At the Austin meeting, SIM scientists discussed a scaled-down version they are calling SIM Lite that would separate its telescopes by 6 metres rather than 9 metres. This would mean a 35% reduction in the weight of the scope, and hence a cheaper launch. Additionally, SIM Lite would dramatically scale down one of its guide interferometers to reduce complexity and save cost. Full cost estimates have not been made yet for SIM Lite, but project scientist Mike Shao of JPL says it might be of the order of \$1 billion.

That, however, may not be enough to make it at NASA. Morse says there may be money available for only a medium-class mission costing \$600 million–\$700 million and, moreover, the agency wants SIM to compete for it along with other proposed missions. Alan Boss, a theorist at the Carnegie Institution of Washington, says a 6-metre SIM Lite "can still do great science". But there are compromises: instead of studying 130 nearby stars for evidence of Earth-like planets, SIM Lite would be able to study only 65, and also gather data at a much lower rate.

For now, it is unclear what will happen to the \$60 million allocated to SIM for the 2008 fiscal year. The language in the congressional bill specifically says it must be used for SIM — it cannot be spent on SIM Lite, Morse explains.



Congress wants to spend more on the Space Interferometry Mission than NASA does.

SIM's story shows how such congressionally directed spending can foul up agencies' best-laid plans, claims James Savage, who studies academic 'pork-barrelling' — the designation of public funds for use in a politician's home district — at the University of Virginia in Charlottesville. The SIM money is not, strictly speaking, an 'earmark', a line item in the budget for an unrequested project. But Savage says a rising tide of scientific lobbying for specific projects has thwarted more general advocacy efforts to boost science funding as a whole.

In the 1980s, there were just a few big research institutions that had Washington offices, Savage says. Now, many hundreds of colleges have offices or hire lobbyists. The money these lobbyists reel in has risen, too. In the 2008 spending bills, there were 2,500 research and development earmarks worth \$4.5 billion, according to an analysis by Kei Koizumi at the American Association for the Advancement of Science.

Griffin was himself effectively lobbying when he talked about the consequences of funding SIM, Boss says, knowing that the astronomy community would not want to sacrifice a mission such as the JWST for SIM. And so, at the Austin conference, SIM supporters responded with their own lobbying — wearing electronic lapel pins that flashed: 'Go SIM Go!' and 'Support SIM!'.

Eric Hand and Alexandra Witze

See Editorial, page 223.

## Stem cells: a national project

Japan is scrambling to harness the promise of Shinya Yamanaka's pioneering work that reprogrammed adult human cells into an embryo-like state. With unprecedented speed, the government is pouring money into developing this home-grown field, some of which will go towards funding a new Yamanaka-headed research centre at Kyoto University.

On 20 November, Yamanaka reported using a relatively cheap and easy technique to reprogramme adult human cells into cells almost indistinguishable from embryonic stem cells. He called these 'induced pluripotent stem cells' (iPS cells) for their ability to differentiate into any of the body's cell types.

Just a week later, Japanese Prime Minister Yasuo Fukuda closed the monthly meeting of the national Council for Science and Technology Policy (CSTP) with a plea to accelerate development of the "revolutionary" method: "I want the CSTP to quickly create an environment in which this science, including clinical research, can move forward smoothly."

By 22 December, the science ministry had laid plans to raise the funding on iPS research from ¥270 million (US\$2.5 million) for 2007, to ¥2.2 billion for the 2008 fiscal year, pledging ¥10 billion over the next 5 years. The health ministry will add close to ¥100 million in the 2008 fiscal year directly to Yamanaka, in addition to ¥410 million for regenerative medicine infrastructure, such as a cell-processing centre.

In December, it was announced that Kyoto University would create a research centre dedicated to iPS, funded by the science ministry. The centre, to be headed by Yamanaka, is expected to open in 2009 and to house 10 principal investigators and 100 researchers.

Japanese researchers keen to get hold of iPS cells can apply to the BioResource Center at the Institute of Physical and Chemical Research (RIKEN) in Tsukuba, north of Tokyo, which will start distributing mouse iPS cells from previous work by Yamanaka in March. But most scientists will want to get hold of the viral vectors that Yamanaka used to introduce the four genes. A virtual consortium whose members will be able to share iPS



### EASY ACCESS

Find news on everything from stem cells to space flight.  
[www.nature.com/news/archive/subject/index.html](http://www.nature.com/news/archive/subject/index.html)

cell information and materials without going through time-consuming material-transfer agreements is planned for the Kyoto University centre.

"If they all agree to recognize each other's technology, they might even be able to share information before publication," says Shin-ichi Nishikawa of RIKEN's Kobe-based Center for Developmental Biology, who is tipped to head the consortium. Nishikawa has already been in touch with organizers of a stem-cell consortium in China, and hopes

**"It's rare for Japan to have such an opportunity."**

that researchers everywhere, especially in the Asia-Pacific region, will be able to work together. "It's rare for Japan to have such an opportunity,"

says Nishikawa. "It should be used to encourage diplomacy."

The science ministry is scurrying to pull the new projects together. ¥1 billion, to be distributed by the Japan Science and Technology Agency, will be available for major iPS research projects from 1 April, but the ministry has yet to decide on research themes. It will start taking applications by the end of March and will pick winning projects soon after.

Such sudden investment is rare for the Japanese government, which usually follows the United States' lead in defining promising scientific fields. In the past, this has led to missed opportunities — most famously leaving Japan a bit-player in the Human Genome Project, even though high-throughput sequencing was first proposed there.

David Cyranoski



Japan is hoping to capitalize on the work that made Shinya Yamanaka an international star.

T. KITAMURA/AFP/GETTY

## SPECIAL REPORT

# Nuclear war: the safety paradox

In the second of a series of articles, **Geoff Brumfiel** looks at whether certain nuclear-weapons technology should be shared.

**W**hen a series of weapons tests announces a new member of the nuclear club, as in both Pakistan and India in May 1998, the natural response is to do everything possible to punish the proliferator and limit its future nuclear development. But some nuclear experts are drawn to the merits of the opposite course of action — supplying advice and technological aid. The argument is that if the world must have more nuclear weapons, it is in everyone's interests that they are safe ones.

To some, the idea of a safe nuclear weapon is the ultimate oxymoron. But the term has fairly clear meanings. A nuclear weapon is at least comparatively safe if it can go off only where and when the government that made it wants it to: not by accident, not on the say-so of a relatively junior officer in the field, and not after it has been stolen by terrorists (or anyone else).

Engineers and scientists in the established nuclear powers have spent decades developing safety mechanisms to ensure that weapons neither explode nor can be exploded if they are involved in accidents or mislaid. This is a real danger; when a B-52 bomber crashed into a tanker plane over Spain in 1966, three of its bombs crashed to earth and one was, for a while, lost at sea.

The details of such safety systems have remained largely classified. But growing instability in Pakistan has created interest in sharing this technology. An article in *The New York Times* in November claimed that US government experts had unsuccessfully pushed for sharing specific safeguard technology, whereas an earlier report by NBC News suggested that some sharing may have already occurred. And at a 5 January debate of Democratic presidential candidates in New Hampshire, Hillary Clinton said that she would advocate that Pakistan work with delegates from the United States and United Kingdom to develop a "fail-safe" for the weapons.

The wisdom of such transfers is hotly

debated among scientists and arms-control experts. "There seems to be a battle," says Michael Levi, a fellow at the Council on Foreign Relations in New York city, "between the lawyers and the technologists." On one side are those who believe such collaboration would undermine the Nuclear Non-Proliferation Treaty (NPT), the keystone of the world's effort to contain nuclear weapons. On the other are those claiming that the dissemination of such technology may ultimately prevent an act of terrorism or an unintended nuclear war.

Regardless of how dangerous a nuclear state may seem, nuclear weapons that are not under the leadership's control are worse, argues Jeffrey Lewis, director of non-proliferation at the New America Foundation, a Washington-based think-tank. "I think there's a need [for the safety technology], and I think it should be shared," he says.

## Kill switch

There are two types of bomb safety device: those that stop a bomb from going off accidentally; and those that stop it from going off without proper authorization. Mechanisms for accident-proofing a bomb range from simple housekeeping (keep the explosive triggers entirely separate from the nuclear cores) to sophisticated design requirements such as 'one-point safety'. In a one-point-safe design, a nuclear explosion will not occur even if one of the various chemical explosive charges in the trigger goes off. This is quite a hard trick to master: before a 1992 voluntary test moratorium, the United States conducted 32 nuclear tests to establish one-point safety on each of its weapons.

Ensuring proper authorization is the role of what America calls a Permissive Action Link, or PAL. PALs are devices that keep the explosive systems of a bomb or warhead isolated from the outside world unless they are unlocked with a specific code: no code, no explosion. If the incorrect code is entered a set number of times, the PAL will disable the weapon, some-

**"As long as you're not teaching them how to improve the function of their warhead, I don't see anything wrong."**



How safe are the nuclear weapons fielded by India (main picture) and Pakistan (inset)?

times with a small explosive charge. After that, the weapon will need extensive servicing before it can be returned to readiness.

Precisely what safety systems various nuclear states have is not open knowledge (the British television news programme *Newsnight* recently caused a stir when it revealed that Britain lacks a PAL system). But their limited system experience and short testing history make it almost certain that any safety systems fielded by new nuclear nations will not be as sophisticated as American ones, says Geoffrey Forden, a physicist and arms-control analyst at the Massachusetts Institute of Technology in Cambridge.

Pakistan, for example, is believed to keep its weapons safe through disassembly, keeping the nuclear cores and triggering explosives in separate locations. But little is known about how the separation is maintained, or how the assembly and arming processes are controlled.

Forden believes that without advanced safety and security systems, such weapons could be co-opted by terrorists or accidentally detonated. Particularly in the case of an accidental explosion at a military base, he argues, "the chances are that they'd think it was an attack",

**HAVE YOUR SAY**

Comment on any of our news stories, online.

[www.nature.com/news](http://www.nature.com/news)



with the Pakistan Army who now teaches at the Naval Postgraduate School in Monterey, California.

Lewis suggests that one solution might be to avoid any active cooperation and simply declassify earlier generations of PAL technology as a resource for other countries. Another possibility would be to educate scientists on the general principles of PAL systems without providing technical details, says Sidney Drell, a physicist at the Stanford Linear Accelerator Center in

California who has examined nuclear-weapons issues.

**"It's not the job of technologists or lawyers. We need a broader, more coherent approach."**

But such collaboration could still fall foul of the NPT (see *Nature* 451, 107; 2008). Article I of the treaty prohibits its assisting non-nuclear-weapons states in the manufacture of nuclear devices. Sharing PAL technology with others outside the NPT could easily be seen as contravening that prohibition, according to Wyn Bowen, head of research in the defence studies department at King's College London. "It's really against the spirit of the treaty," he says.

Joseph Cirincione, director for nuclear policy at the Center for American Progress, a Washington-based think-tank, concurs. "The solution is not to build bombs with better controls," he says, "but to eliminate the bombs we have."

Lewis counters that PALs would not alter the yield or military purpose of a weapon: "As long as you're not teaching them how to improve the function of their warhead, I don't see anything wrong with that."

For den sees further technologies as ripe for sharing — for example, a global network of early-warning satellites that could provide all nations with information about missile launches in not-quite-real time. Access to such a system would not give countries early warning of real attacks. But after an unexplained blast or accident, the system would allow the nation affected to see whether there had been any hostile launches that might explain the blast. Coyle is sceptical about whether countries could be persuaded to use such a system, however.

In many ways the real problem is that there are convincing arguments on both sides, Levi says. "You have two narrowly focused camps," he says. "But frankly, it's not the job of technologists or lawyers. We need a broader, more coherent approach." Ultimately, he argues, that approach can come only from politicians. ■



and would retaliate with nuclear force.

But sharing the details of PALs and other safety systems raises a range of problems. For one thing, sharing details about implementing the technology would also mean exchanging some information about the weapons for which it was developed. PALs must be placed at a critical point in the design, says Philip Coyle, a former designer now at the Center for Defense Information, a Washington-based defence think-tank. Sharing the location of PALs and the mechanism by which they work would be "on the edge of possibly revealing information about the design", Coyle says. There is thus the risk that new nuclear-weapons states could learn at least some details about US

nuclear weapons, and accordingly improve the capabilities of their own designs.

Another concern is that the safer the weapons become, the more comfortable a state such as Pakistan might feel about deploying them on the front line. This could negate, or even reverse, any advantages resulting from the improved inherent safety of the weapons themselves.

From the recipients' point of view, getting such technology means giving scientists from another country at least some details of bomb design. "No way will Pakistan be sharing with the United States or any other country any data or information about its nuclear programme," says Feroz Khan, a former brigadier general

K. KISHORE/REUTERS

M. KHURSHED/REUTERS

## ON THE RECORD

## “Requires belief.”

Footnote to a statement about the placebo effect on the FairDeal Homeopathy website.

## NUMBER CRUNCH

**1.8 metres** is the average height of Dutch men, thought to be the world's tallest people, since 2001.

**0.03 metres** is the average increase in the height of Dutch men from the 1980s to 2000.

**0 metres** is the change in average height since 2001, leading researchers to conclude that the Dutch have stopped growing.

## SCORECARD

**Carrots**

Patients with type 2 diabetes who increase their sugar intake through carrot cake show no adverse changes in blood sugar, provided that they don't put on weight.

**Dull vegetables**

Encouraging farmers to grow shiny crops instead of matt ones could reduce a region's maximum daytime temperature by 1.9 °C and fight global warming.

## ZOO NEWS

**Avoiding Knut-mania**

After declaring that it would avoid the media's obsession with Knut and allow its own newborn polar bear cubs to starve, Germany's Nuremberg Zoo has had a radical change of heart. It has helped set up a website dedicated to cute images of the surviving four-week-old cub, inviting the public to name the little star. The site is getting 15 name suggestions a minute, leading Sidelines to wonder if the zoo's daily news conferences on the cub will be enough to satisfy its fanbase.

Sources: [www.fdhom.co.uk](http://www.fdhom.co.uk), AP, The Sugar Bureau, The Guardian

# Europe to capture carbon

New power stations across Europe could be routinely fitted with carbon-dioxide capture and storage (CCS) technology within two years under a proposal by the European Commission.

Next week, the commission will propose a directive on geological storage of CO<sub>2</sub> that would require all new fossil-fuel combustion plants to have “suitable space on the installation site for the equipment necessary to capture and compress CO<sub>2</sub>”. Builders of new plants would need to assess the availability of “suitable storage sites and the technical feasibility of CCS retrofit” before being granted construction licences. If the European Parliament and Council approve the proposal, it could become law in the European Union's 27 member states as early as 2009.

Champions of CCS applaud the move as a milestone. “We would have hoped for a specific date for CCS to become mandatory,” says Paal Frisvold, chair of Bellona Europa, a Norway-based environmental group. “But even so, we do think that the proposed directive is an absolutely crucial step towards making industrial-scale CCS a reality.”

Experts think that CCS could reduce global CO<sub>2</sub> emissions by one-third by 2050, if widely deployed. The proposal is in line with the European Commission's Strategic Energy Technology Plan, released last November, which prioritized development and commercial deployment of CCS to reduce emissions in Europe.

The proposed directive is the first attempt anywhere in the world to provide a comprehensive legal framework for industrial CCS activities, from storage-site selection, to environmental monitoring, to liability issues. And

it ensures that CO<sub>2</sub> captured and stored will be credited as not emitted under the European Union's mandatory emissions-trading scheme.

Economic incentives are vital to getting industry onboard, the European Commission believes. It hopes that if the costs of capturing CO<sub>2</sub> become lower than the costs of releasing the gas into the atmosphere, industry will voluntarily switch to CCS technologies.

But the technology is not yet mature. Scrubbing CO<sub>2</sub> from the gas stream is expensive and decreases the efficiency of coal-fired plants, so it is not yet commercially viable.

In a white paper also to be released next week, the commission will spell out measures

**“The proposed directive is a crucial step towards making industrial-scale CCS a reality.”**

and incentives for ‘early movers’ in the power industry. Among other things, financial aid worth up to €1.5 billion (US\$2.2 billion) could be provided to encourage industry to set up 10–12 demonstration facilities in the next decade. However,

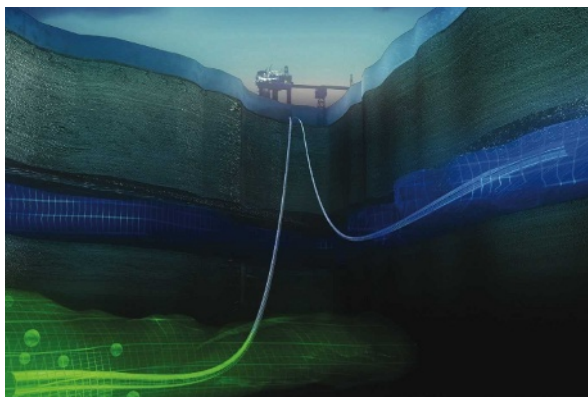
commission officials say they doubt that more than four or five demonstration plants will end up being built.

At the moment, in Europe, only Norway and Britain have concrete plans for pilot plants. The Norwegian company Statoil currently operates Europe's only commercial CCS project, in the Sleipner West natural-gas field in the North Sea. “There is still resistance in the power industry, but the amount of scepticism is waning compared to five years ago,” says Frederic Hauge, vice-chair of the European Union's Technology Platform for Zero Emission Fossil Fuel Power Plants, which was established by the European Commission in 2005 and comprises scientists, industry and non-governmental organizations.

Meanwhile, China and the United States also

plan to build large-scale demonstration plants in the next 10 years. An Illinois-based site for FutureGen, a \$1.5-billion public-private partnership to build a coal-fuelled near-zero-emissions power plant, was announced in December. This January, America began testing the safety, permanence and economic feasibility of storing large volumes of CO<sub>2</sub> in geological structures at 22 test sites run by 7 regional partnerships, each comprising universities, state agencies and private companies.

Quirin Schiermeier



Europe has the world's only commercially viable carbon capture and storage facility — at a gas field in the North Sea.

ALLIGATOR FILM/BUG

**THE GREAT BEYOND**

Our news blog digests what is being reported elsewhere.  
<http://blogs.nature.com/news/thegreatbeyond>

# Novartis Foundation to close its doors

Having given nearly 60 years of intellectual succour and hospitality to biomedical scientists from around the world, London's Novartis Foundation will close at the end of February. Its historic building will probably be taken over by the Academy of Medical Sciences charity, which is unlikely to be able to afford to continue the foundation's tradition of intimate symposia.

Established in 1949 as a scientific meeting house by the Swiss drug company Ciba-Geigy, the foundation launched its series of intensive three-day symposia the following year. The formula of the symposia, with their extensive discussions and accompanying open discussion meetings, was



The Novartis Foundation is widely renowned for its symposia.

similar to that of the Gordon Research Conferences, based in West Kingston, Rhode Island. The foundation has run more than 400 symposia, as well as a publishing programme and other activities. Between conferences, any scientist visiting London from around the world could stay at the foundation, chat with other guests over the huge, maple-wood breakfast table and enjoy the library and lounge.

In 1996, Ciba-Geigy was merged into the pharmaceutical giant Novartis, which moved its research headquarters from Basel to Boston, Massachusetts, in 2002. Soon after, Novartis decided that the foundation was no longer relevant to its interests. "The meetings did not allow us to

maximize our impact," says a company spokesman. Quiet attempts by the directors and trustees of the foundation to find a new corporate sponsor failed. Negotiations are now being completed for the transfer of the premises to the burgeoning Academy of Medical Sciences, which celebrates its tenth anniversary this year.

In a statement, the academy says that it would like to "build on the scholarly tradition" of the foundation and retain the reputation of its building as a "hub for scientific exchange and networking". But the academy is strapped for cash and a continuation of the international meetings is thought unlikely.

Scientists will miss the institution sorely. "It has been an academic haven in the centre of London," says neuroscientist Colin Blakemore of the University of Oxford, UK, a member of the foundation's executive council. ■

**Alison Abbott**

See Editorial, page 224.



R. RESSMEYER/CORBIS

China is the first country to begin a project to sequence the whole genomes of large numbers of private individuals.

## Genomics sizes up

Next-generation human genomics has arrived. The first large-scale whole-genome sequencing project has now begun in China, and an international multi-genome sequencing programme is hot on its heels.

The Yanhuang Project, which will sequence the entire genomes of 100 Chinese individuals over 3 years was announced by the Beijing Genomics Institute (BGI) on 8 January. Ye Jia, a spokeswoman for the project, said that once it is completed, the BGI aims to sequence the genomes of thousands more people, including ethnic groups from other Asian countries.

And a large international project, which aims to sequence the genomes of close to 1,000 individuals, is expected to be formally unveiled by the US National Institutes of Health in Bethesda, Maryland, and the Wellcome Trust Sanger Institute in Cambridge, UK, later this week. As yet it doesn't have a name, but is informally called the '1,000 genomes' project and the 'Multigenome project'. It will probably include the hundreds of individuals who participated in the International HapMap Project — an ongoing study of genetic diversity — as well as hundreds of other individuals.

The BGI will also participate in the 1,000 genomes project, says director Yang Huanming. However, only participants who meet the ethics and consent rules decided on by the international collaboration will be able to join that study, he says.

The projects usher in what many scientists think will be a new era of large-scale genomics

— made possible with rapid-sequencing technologies — that will lead to more powerful comparisons between and within populations. Last year, scientists Craig Venter and James Watson became the first to release their complete individual DNA sequences. And a team led by George Church at Harvard University in Cambridge, Massachusetts, has begun the 'Personal Genome Project' that will examine portions of DNA from ten individuals who have agreed to share their information with the rest of the world.

But the Yanhuang Project — named after two emperors thought to be the ancestors of China's largest ethnic group — is the first to examine the entire genomes of private individuals. The first individual sequenced in the Yanhuang Project was a researcher; the second paid 10 million yuan (about US\$1.4 million) to have his genome sequenced, Yang says. It is unclear whether such people will qualify for the international project, whose rules on confidentiality of data and the informed consent of participants may differ from China's.

Whole-genome sequencing studies are expected to deepen our scientific understanding of populations such as the Chinese, whose genetics have not been studied in great detail. The findings will inform medical research specific to those populations, and improve our understanding of human history, says Rasmus Nielsen of the University of California,

Berkeley. "One of the exciting things about having so many sequences from Chinese individuals is that we will be able to say how much genetic exchange there has been between continents since [early humans migrated] out of Africa. That's been very hotly debated."

The sequencing will allow scientists to add more detail to their maps of human diversity. The last large study of diversity, the HapMap, analysed only single-nucleotide polymorphisms, or SNPs — places in

which DNA differs between two individuals by just one letter of the genetic code. This approach allows scientists to hunt for relatively common genetic variants.

But the evidence linking disease to rare variants is growing, says Richard Myers, director of the Stanford Human Genome Center in Palo Alto, California. Whole-genome sequencing will improve detection of these rare variants, and offer a more complete understanding of the genetics of many human traits, he predicts.

"It's going to be very useful to sequence genomes from all populations and have large enough numbers so you can do comparisons between populations," Myers says. "Even if you don't care about disease, it's going to help us look at human population history and phenotypes not relevant to disease, such as craniofacial structure, eye colour, hair colour and other fascinating things."

Jane Qiu and Erika Check Hayden

**"It's going to be very useful to sequence genomes from all populations."**

## Nuclear power gets green light from UK government

The UK government is endorsing the construction of nuclear power plants to help reduce greenhouse-gas emissions. In a white paper released on 10 January, the government promised to streamline licensing procedures, citing global warming and energy security as the driving factors.

The announcement was hailed by supporters of nuclear power as a major step towards an increase in nuclear capacity. "It's a very robust move forward," says David King, the government's former science adviser who is now at the University of Oxford. But environmentalists say that the decision will do little, if anything, to reduce Britain's greenhouse-gas emissions, the vast majority of which come from natural gas and oil use.

Shortly after the announcement, EDF, a French-based firm that is the world's largest operator of nuclear plants, said that it hoped to build up to four reactors on existing nuclear-power sites in Britain.

## Health agency recalculates death toll for Iraq conflict

A survey by the World Health Organization (WHO) has estimated the violence-related death toll in Iraq, between 2003 and 2006, at 104,000–223,000 (Iraq Family Health Survey Study Group *N. Engl. J. Med.* 358, 484–493; 2008).

The figure is higher than the 47,000 figure given for the same period by the Iraq Body Count, an organization that bases its tally mainly on media reports. And it is much lower than the controversial 426,400–793,700 deaths estimated by

researchers from Johns Hopkins University in Baltimore, Maryland, and the School of Medicine at Al Mustansiriyah University in Baghdad, Iraq (see *Nature* 446, 6–7; 2007).

The latest survey involved a large team of officials from the WHO and Iraq and covered 9,345 households, compared with 1,850 in the Johns Hopkins study. Some critics say that the sample was still too small, but others say that given the difficult conditions in Iraq, and the robustness of the methodology, enough data have been gathered to make the estimated death toll plausible.

Team member Mohamed Ali notes that "nearly 200,000 deaths is not a small number".

## Florida funds expansion of Oregon university

In an unprecedented move, Florida has lured a public university in Oregon to the sunshine state with an offer of \$118 million to establish a research laboratory there.

Oregon Health & Science University in Portland last week announced that its Vaccine & Gene Therapy Institute (VGTI) in Beaverton would expand to Port St Lucie in Florida — where government officials are aggressively funding research facilities to spawn a biotechnology industry (see *Nature* 442, 729; 2006). It is thought to be the first time that one US state has paid for biomedical research facilities for another state university.

The VGTI, which currently has about 90 staff working for seven principal investigators, expects its facility in Florida to be more than double the size of its Oregon site, which will continue to operate. Florida is providing \$60 million for operations over a decade, and local governments will fund infrastructure costs.



Stanford's B-meson work is coming to an early end.

SLAC

## Budget cuts force early closure of Stanford collider

In early March, California's Stanford Linear Accelerator Center (SLAC) will shut down a collider that produces B mesons. The closure means that the lab's commitment to BaBar — an international collaboration studying the differences between matter and antimatter — will now end seven months early.

The announcement was made on 7 January by SLAC director Persis Drell after the US Department of Energy gave her its plan to deal with deep budget cuts to high-energy physics. Faced with a choice between keeping SLAC's 'B-factory' open and continuing to run the Tevatron, the high-energy collider at Fermilab in Batavia, Illinois, the department chose the Tevatron, which might detect the Higgs boson before the Large Hadron Collider is turned on at CERN, Europe's particle-physics lab based near Geneva, later this year.

SLAC also plans to lay off 125 of its 1,600 employees in April, on top of an ongoing 100-person reduction.

## Time is running out for paranormal prize

Challengers for the US\$1-million prize offered by the James Randi Educational Foundation for proving paranormal powers have just over two years left to claim the cash. Randi has announced that the paranormal-activity challenge, in which contestants must demonstrate their powers 'under proper observing conditions', will end on 6 March 2010 — exactly 12 years after he first offered up the prize money.

Randi says that the challenge was intended to tempt high-profile paranormal-activity celebrities to come forward. In 2007, Randi changed the rules of the prize so that applicants were only eligible to enter if they had a media profile and some form of academic endorsement. But as the prize remains unclaimed, and the highest-profile celebrities have not entered, Randi would rather the million dollars were freed to be used elsewhere in his foundation, he says.

## Free bags face the axe in China

China is clamping down on plastic shopping bags in a bid to clean up the environment and save energy.

From 1 June, shopkeepers will no longer be allowed to hand out plastic bags to their customers for free. Failure to charge for the bags could result in a fine. And the manufacture and sale of 'ultrathin' bags — less than 0.025 millimetres thick — will be banned from the same date.

Although this should be good news for the environment, customers feel they are being unfairly burdened. A poll of consumers by the *People's Daily*, the official communist newspaper, showed that more than half opposed the ban.

South Africa, Ireland and Bangladesh have already banned or taxed plastic shopping bags and other countries, such as Australia, are considering following suit.



TEH ENG KOON/AFP/GETTY



# COSMOS IN A BOTTLE

Physicists often borrow techniques from other fields. But how far can this get you? **Geoff Brumfiel** asks if simple table-top experiments can provide new insights into the early Universe.

**T**ake a look at water running in a sink and you'll see an intriguing everyday phenomenon. As water from the faucet strikes the basin, it will create a small saucer of moving water. The water entering this saucer from above flows smoothly and radially out; its even flow creates a ring of ripples which holds the more turbulent water in the rest of the sink at bay. Outside the ring, the water is full of waves and eddies, but on the inside, the water is moving out too fast for the ripples to penetrate — no information from the rest of

the sink can cross into the circle.

One of the long term goals of the astronomical community is to produce images of the 'event horizons' that surround black holes — the ultimate points, or rather surfaces, of no return. Theoretical physicists have spent decades calculating what happens at event horizons, and astronomers now want to spend decades more, and billions of dollars, trying to see what one actually looks like. However, other physicists think that they can get at least some of the answers to that question by

studying those rippling fluid rims in the sink.

The analogy between sink-saucer and black hole isn't perfect. For one thing, water flows out from the horizon line into the sink, while quite the reverse happens in a black hole.

But according to Bill Unruh, a theoretical physicist at the University of British Columbia in Vancouver, Canada, it is closer than you might think. In the early 1980s Unruh imagined a similar sort of flow as a thought experiment: a waterfall in which the falling water exceeded the speed at which sound waves could travel in the fluid<sup>1</sup>. In that system it is the point when water reaches the speed of sound that creates an 'event horizon' beyond which sound can never escape. "If you set up the flow right," he says, "you could exactly mimic a black hole."

D. ALLISON

## Getting the flow right

Since that time, a small coterie of physicists has devoted itself to simulations of esoteric phenomena such as black holes and the workings of the early Universe. But before anyone starts to think about saving billions of space-faring dollars with some cleverness in the kitchen sink, there are a few caveats. Getting "the flow right", as Unruh puts it, tends to mean using superfluid liquid helium only a fraction of a degree above absolute zero, or some even more esoteric system, such as a set of ultracool trapped atoms in a Bose-Einstein condensate — another close-to-absolute-zero fluid with quantum properties. Most of the proposed setups haven't even made it off the drawing board; only a handful of experiments have been successfully carried out.

And then there's the problem of what, if anything, such models actually tell you. If system B mimics system A in a set number of ways, and goes on to exhibit some other hitherto unlooked for activity, does that mean that system A does the same thing? Or does it mean that the two systems are not that similar after all?

Despite these worries, kitchen-sink or table-top cosmology continues to generate excitement among a small but fervent group of physicists, mostly in Europe, where there is a small but steady stream of funding for such research. Much of the work involves superfluid helium, a good medium for studying phase transitions — transitions from one state to another — and quantum effects, both subjects of great importance in cosmology. Later this month, those interested in condensed matter

and cosmology will gather at the Royal Society in London to discuss the future of their attempts to mimic — and manipulate — the otherwise unobservable. “You’re never going to do experiments *in situ*,” says Tanmay Vachaspati, a cosmologist at Case Western Reserve University in Cleveland, Ohio. “It has to be in a laboratory setting.”

### Cosmic inflation

The field of condensed matter, which covers everything from waterfalls to semiconductors, has always been a useful source of inspiration for those interested in the origin of the cosmos, according to Paul Steinhardt, a cosmologist at Princeton University in New Jersey. In the mid-1980s, he was working on refining a theory known as cosmic inflation that postulates that the Universe underwent a period of extremely rapid expansion shortly after the Big Bang. The problem at the time, Steinhardt says, is that nobody knew how to explain how the transition from inflation to today’s more slowly expanding Universe occurred. The dominant thinking then was that the present day Universe would have begun as bubbles in the inflationary cosmos. But the bubbles, according to calculations, would be nothing but vacuums — matter and energy would never have developed under such conditions.

Steinhardt himself was stuck until he read a description of unusual ‘phase transitions’ in a mixture of helium isotopes. Normal fluids change their phase — from gas to liquid, say — following a bubble regime similar to the one that theorists believed ended inflation. But the mixture of superfluid helium changed its properties in a completely smooth, uniform fashion. Applied to cosmology, the superfluid transition allowed the entire Universe to gently roll from inflation to the present-day conditions, says Steinhardt.

Since Steinhardt’s work, superfluid helium has emerged as the material of choice in these sorts of experiments. In particular, helium-3, an isotope of helium with two protons and one neutron, has very unusual properties, which make it an unusually good proxy for the cosmos.

In addition to exotic phase transitions, helium-3 can undergo the phenomenon of ‘symmetry breaking’. Normally, pairs of atoms in the liquid have their spin and orbital angular momentums aligned in random directions. But when cooled, the helium atoms will snap into a single alignment. The process is somewhat like iron filings lining up in a magnetic field, except that the helium arranges itself

spontaneously — creating order from chaos. Physicists believe that symmetry breaking in the early Universe led to the creation of every force except gravity.

Taken together, the symmetries and phases of a helium-3 superfluid give the quantum liquid an important Universe-like quality, says Grisha Volovik, a condensed-matter theorist at the Helsinki University of Technology in Finland. “All the ingredients are certainly there,” he says.

So how far can such analogies can be trusted? And what if the cosmological theories being tested are themselves wrong? Around the time at which Steinhardt was refining his inflation theory, a theoretical physicist at Imperial College in London, Tom Kibble, was working on an alternative model. Kibble had a theory that the cooling of the early Universe as it expanded also created massive structural defects — called cosmic strings — that were the seeds of the large network of galaxies we see today.

Kibble’s hypothesis worked perfectly in helium-3, where rapid cooling led to a tangle of ‘quantum vortices’ that matched his theory. Unfortunately, he says, his cosmic strings theory of galactic structure failed to match up with astronomical observations of the cosmic background radiation left over from the Big Bang. After satellites designed to study the cosmic background delivered their results in the early 1990s, Kibble says: “It became clear that the predictions of inflation were rather good, and the predictions of cosmic strings were completely wrong.”

In other words, laboratory models had verified the theorists’ equations, but they had provided

absolutely no insight into whether those equations could be applied to the cosmos.

That early failure left many experimentalists and theorists sceptical of any bench-top models of the early Universe. “Frankly,” says Wolfgang Ketterle, a Nobel-Prize-winning condensed-matter physicist at the Massachusetts Institute of Technology in Cambridge, “I don’t think a table-top experiment will answer fundamental questions about the cosmos any time soon.”

### Stringing it together

Such concerns have not stopped Richard Haley of Lancaster University, UK, from pursuing lab analogues for string theory — perhaps the most experimentally intractable

theory of fundamental physics. String theory is controversial because it has evolved over the past two decades almost without reference to experiments or observations, and so some critics view it more as a branch of mathematics than of physics.

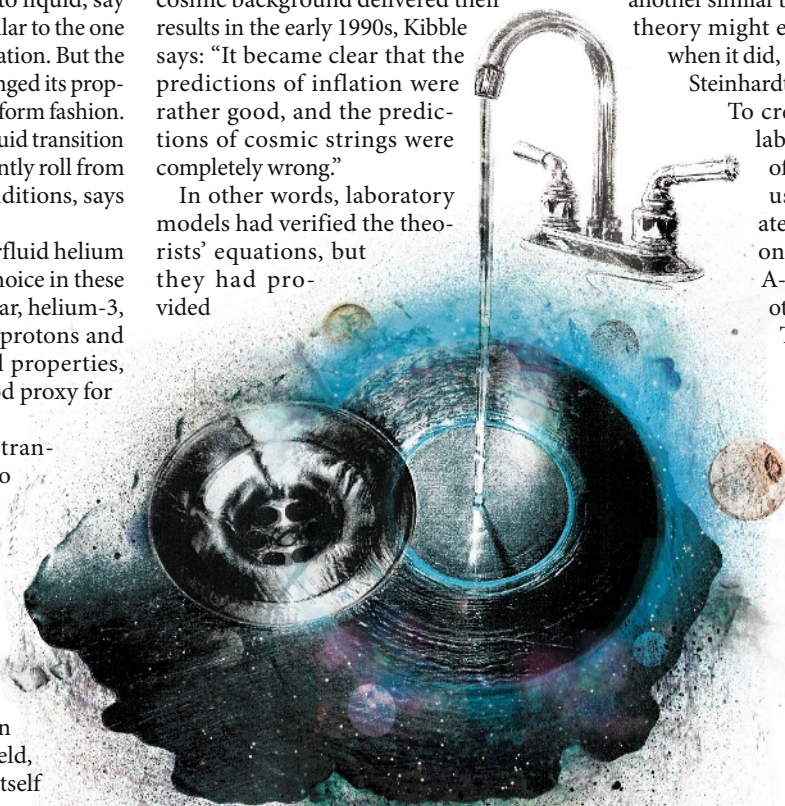
Some versions of string theory postulate that our Universe may sit on a three-dimensional membrane, or ‘brane’, suspended in a higher-dimensional space, the way a two-dimensional sheet of paper sits in the three-dimensional world. In such models, string theory explains the end of the inflationary period through the collision of our brane with another similar brane. If it were true, the brane theory might explain why inflation ended when it did, a question left unanswered by Steinhardt’s earlier work.

To create colliding branes in the lab, Haley brought two phases of helium-3 together. His team used a magnetic field to create a helium-3 sandwich, with one part of the superfluid, the A-phase, as the filling and the other, the B-phase, as the bread. They then decreased the field strength and watched as the two B-phases collided<sup>2</sup>. Mathematically speaking, Haley says, the phases are good analogies for cosmic branes.

In Haley’s experiment the colliding phases did not merge smoothly into one uniform B-phase, but instead left behind structural defects — most likely quantum vortices of the same sort predicted by Kibble. If these swirling vortices have analogies in the Universe,

**“If you set up the flow right, you could exactly mimic a black hole.”**

— Bill Unruh



G. BRUNFIELD/D. ALLISON

then they should be detectable as massive cosmic strings. Unlike Kibble's original idea, these strings would be a smaller fraction of the Universe's mass, but they should still be detectable by using ground and space-based interferometers to observe gravitational waves. Haley, meanwhile, says that he and his team are now working to further understand the different kinds of vortices created by the collision.

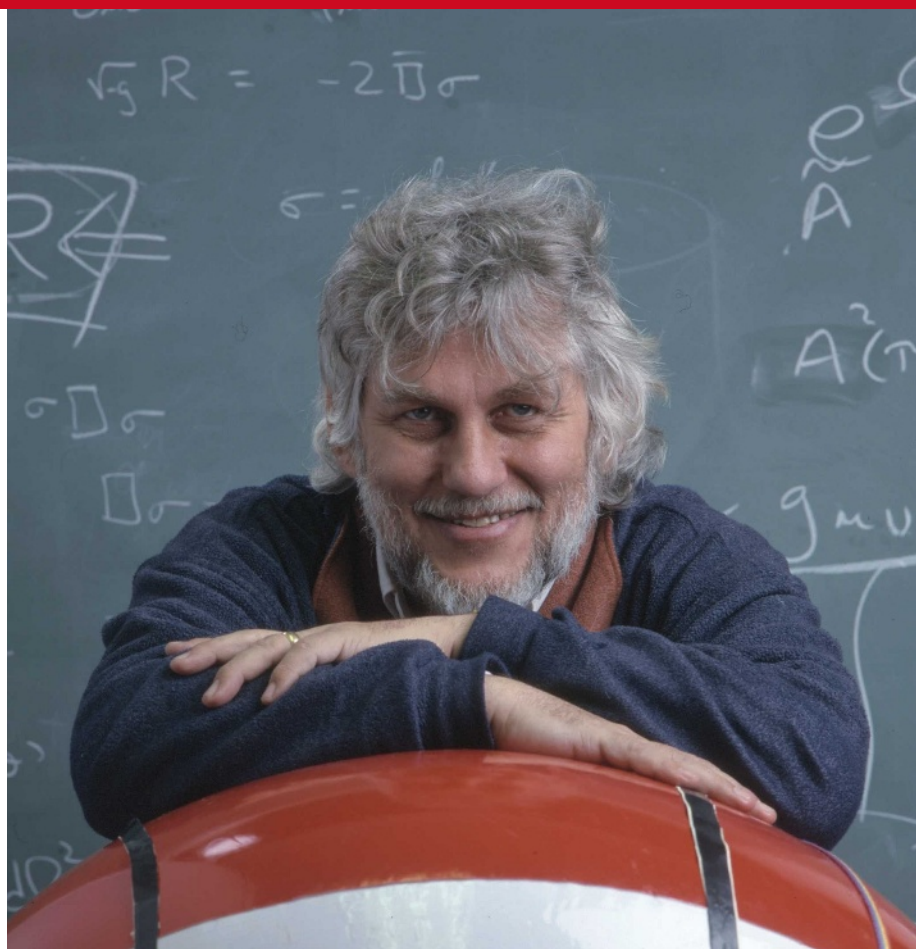
### Testing the untestable

Of course, cosmologists need to apply caution when interpreting such lab-based results. Steinhardt notes that string branes are flat and attract one another, whereas the helium-3 'branes' are curved and have no attractive force. The model is far from perfect. Still, in a field such as string theory where exotic mathematics reigns supreme, an experiment that makes any testable prediction could have a big impact, says Joe Polchinski, a string theorist at the Kavli Institute for Theoretical Physics in Santa Barbara, California. "You never know what you might find," he says.

From the experimentalist's perspective, even a failed analogy can find another purpose. The quantum vortices first predicted by Kibble and his colleague Wojciech Zurek, a quantum theorist at the Los Alamos National Laboratory in New Mexico, are now being used to track the movement of helium-3 in other experiments, according to Matti Krusius of Helsinki University of Technology. "This is a nice phenomenon," he says. "We use it to study turbulence."

Such crossover from cosmology into condensed matter is a common and overlooked benefit of these collaborations, says Ralf Schützhold, a quantum theorist at the Technical University of Dresden in Germany. Because the Universe has been expanding since the time of the Big Bang, cosmologists' equations that model this expansion can work well for systems that are changing. That makes them particularly useful for understanding phase transitions and other phenomenon. "It's very nice to consider effects in condensed matter based on these beautiful equations from cosmology," he says.

Schützhold and his team are now working on a different cosmological analogue that could help to explain the origin of matter and energy in the Universe. Under normal circumstances, atoms are constantly moving, but when a single atom is chilled to near absolute zero, its real motion converts into 'virtual' quantum fluctuations, which are temporary changes in the amount of energy in a small volume of space. Following inflation, cosmologists believe that the Universe underwent the reverse of that process: virtual quantum



Bill Unruh hopes that experiments will mimic the behaviour of a black hole in the laboratory.

fluctuations in the vacuum of space became real matter and energy. A laboratory experiment on a single atom, Schützhold says, could allow him and others to see how thermal noise and other real-world effects altered the fluctuations that created the cosmos we see today.

### Good vibrations

Controlling a single atom is no small task, but Tobias Schätz, Schützhold's experimental partner at the Max Planck Institute for Quantum Optics in Garching, Germany, says that he is reasonably confident that it can be made to work. Even if it can't, he says, the project is likely to aid his work in quantum computing. "We have to work in this direction anyway," says Schätz.

That's just as well, because experiments to realize the quantum vibrations of an atom require exquisite control of the laser system used to cool it. "It is really pushing experimental technique to its limits," says Ketterle. Basing a career on such analogies would be "scientific suicide," he says, especially given their tentative link to actual cosmology.

Experiments on the black-

hole models that Unruh first described are even further off. Efforts to create a waterfall equivalent in helium-3 have been stymied by fluid turbulence. Other approaches are now in the works: some groups are working with Bose-Einstein condensates<sup>3</sup>, which can be studied at lower flow speeds than helium-3. Other techniques employ a series of light pulses in special fibre-optic cables<sup>4</sup>.

Ultimately, a lab analogue that displays quantum behaviour is needed. Such a system could allow experimentalists to observe Hawking radiation — a quantum-mechanically induced glow that theorists predict exists around the event horizon. The pay-off for theorists in this case promises to be tangible: an observation of Hawking radiation in such a system could inform debate about whether and how black holes 'evaporate' over time.

So despite nearly two decades of waiting for his black-hole analogy to reach fruition, Unruh's enthusiasm for the project remains undimmed. "It's a really neat idea and it would be great if it works," he says, then adds: "I'm astonished every time I see what these experimentalists can do."

**Geoff Brumfiel is a senior reporter for Nature based in London.**



**"I don't think a tabletop experiment will answer fundamental questions about the cosmos any time soon."**

— Wolfgang Ketterle

1. Unruh, W. G. *Phys. Rev. Lett.* **46**, 1351-1353 (1981).
2. Bradley, D. I. et al. *Nature Phys.* **4**, 46-49 (2008).
3. Garay, L. J., Anglin, J. R., Cirac, J. I. & Zoller, P. *Phys. Rev. Lett.* **85**, 4643-4647 (2000).
4. Philbin, T. G. et al. arXiv:0711.4797v1 (2007).



# POWER PLAY

A German physicist and a hedge-fund magnate are competing to push protein simulations into the realm of the millisecond. **Brendan Borrell** finds out what is at stake.

For a while, Klaus Schulten did not mind the Godiva chocolates arriving in his team's mailboxes at the University of Illinois in Urbana-Champaign. Nor was Schulten, whose biophysics group boasted one of the fastest algorithms for simulating protein structures, much concerned when his programmers received e-mails heralding a job opportunity at an undisclosed Manhattan firm that aimed to "fundamentally transform the process of drug discovery".

It was early 2004, and Schulten's 40-strong group was attracting close to \$2 million a year in grant money. Nearly 20,000 users had downloaded his software, called NAMD for Nanoscale Molecular Dynamics, for use on computers running hundreds of parallel

microprocessors to simulate how individual atoms behave in proteins and other large molecules. Schulten's group itself was working on a million-atom model of the satellite tobacco mosaic virus, which the researchers called "the first all-atom simulation of an entire life form"<sup>1</sup>.

But the German-born physicist got his wake-up call in 2006, when he saw a table of computing benchmarks in a report from that year's supercomputing conference in Tampa, Florida. A new program called Desmond, he saw, could calculate each step of a standard molecular-dynamics simulation — the 23,558 atoms in a system involving the protein dihydrofolate reductase — in a little over a thousandth of a second. NAMD was ten times slower. "Suddenly,"

Schulten says, "we were not the best anymore."

The title had passed to the sender of the chocolates — David Shaw, a hedge-fund magnate and computer expert who taught himself physical chemistry. Over the previous few years, he had recruited more than 50 scientists and engineers, including three former students from Schulten's group, and put them to work in his midtown Manhattan high-rise.

In the paper from the supercomputing conference, Shaw's team wrote that Desmond "is faster than NAMD at all levels of parallelism examined"<sup>2</sup>. And the group noted that on one simulation Desmond ran faster on 1,024 processors than NAMD ran on the 16,384 processors of IBM's Blue Gene/L — the world's fastest supercomputer.

THOMPSON-MCCLELLAN



**Model behaviour:** Klaus Schulten is pursuing his dream of creating a 'computational microscope' to study complex molecular dynamics.

The numbers shocked Schulten, who believed his team was on course to simulate molecular dynamics on the scale of milliseconds — longer than anyone had previously achieved. Even with cutting-edge programs such as Desmond and NAMD, scientists have been able to glimpse only the fastest-folding proteins, such as the villin headpiece, which folds in about 10 microseconds. The number of possible configurations of atoms in larger molecules, over time and in three dimensions, is astronomical. If these kinds of simulation could be sped up 1,000-fold, which even then could take a month of computing time, the pay-off could be high. They might, for instance, reveal binding sites for new drugs to tackle a wide range of medical problems.

Shaw and Schulten are now spending millions of dollars each to break the millisecond barrier. But some in the field aren't sure what the all-out push will come to. As Ross Walker, a computational biologist at the San Diego Supercomputing Center in California, puts it: "A lot of what they are going to see are limitations on the underlying computational models."

### Pushing the envelope

To make molecular-dynamics simulations feasible with today's computers, scientists have had to make a number of simplifying assumptions. Typical simulations calculate the forces acting on each atom from a century's worth of chemistry experiments on organic molecules much smaller than the proteins scientists wish to simulate. The simulated molecules are also pegged together like Tinkertoys; they can change shape during the simulation, but cannot react to form new molecules.

The first software that sought to capture this world was developed at Harvard University in the late 1970s. In a paper in *Nature*, a team led by Martin Karplus published its 458-atom simulation of a tiny protein on an IBM 370, a top-of-the-line supercomputer<sup>3</sup>. Today, development teams around the world continue to work on CHARMM, or Chemistry at Harvard Molecular Mechanics, even as other algorithms such as NAMD have risen to compete with it.

One of the biggest factors limiting the development of molecular dynamics has always been computational power — which is where Shaw comes in. Having stepped back from running his hedge fund around 2001 (see 'From science to finance and back again', overleaf), Shaw, who is also an adjunct professor of biomedical informatics at Columbia University in New York, returned to his first enthusiasm — the architecture of massively parallel supercomputers. Predicting the motions of large systems of atoms requires finding the best way to communicate particle positions and forces among multiple processors. And on a scorching afternoon in June 2003, Shaw holed himself up at a friend's house and found a way to speed things up.

In traditional parallel approaches, each processor calculates forces to update the position of all the particles in its own small box of simulated space. But to do so, it must import positional data from neighbouring boxes within a certain radius. Shaw's strategy, implemented in Desmond, changes the geometry of this import region from a hemisphere to a semi-circular plate and a rectangular tower. As the number of processors available to Desmond

grows, the volume of this import region shrinks more quickly than in the approaches used by NAMD and CHARMM. In one of the first studies to use Desmond, this speed-up gave Shaw and his collaborators an unprecedented view of the workings of an ion transporter that the bacterium *Escherichia coli* uses to maintain its salt and pH balance<sup>4</sup>.

But Shaw knew that software alone could not obtain millisecond-long molecular simulations. His plan has been to build a supercomputer so dumb, he says, that it can do nothing except molecular dynamics. "But," he beams, "it's really fast at that." He calls it a computational microscope and has named it Anton, after Anton von Leeuwenhoek, the seventeenth-century Dutch scientist and builder of microscopes. The first segment of Anton is due to arrive in Shaw's lab at the end of the year.

### Need for speed

Anton uses a high-speed task pipeline to accelerate the most computationally intensive tasks of molecular dynamics — modelling certain long-range interactions among atoms. But the chip does not have the ability to speed up software-based operations to the same extent, and the hard-wired pipeline may not be flexible enough to efficiently incorporate advances in the field. "At this point, though, we placed our bets," Shaw says.

When Shaw began the work, he estimated that Anton would run molecular-dynamics simulations 1,000 times faster than previous parallel supercomputers. In recent months, he has stopped presenting the 1,000-fold estimate in talks, although he still believes Anton will run more than 100 times faster than today's machines. But with general-purpose hardware doubling in speed about every two years,

many wonder how long Anton might maintain a lead. "If you are a little bit of a sceptic," says Schulten, "you would say it is another attempt for a special-purpose processor that will be overrun by market forces."

The field is littered with what Gregory Voth, a computational chemist at the University of Utah in Salt Lake City, calls

"dead bodies". In 1984, the late biochemist Cyrus Levinthal designed a molecular-dynamics computer called FASTRUN, but it took his group six years to get it running. During the past ten years, IBM and RIKEN, Japan's main research institute, have collaborated on several generations of chips intended for molecular-dynamics simulations, called MD-GRAPE, without producing any major breakthroughs in the field. At the National Institutes of Health in Bethesda,

**"I wouldn't have told them about a great solution I had developed, and they wouldn't tell me their solution."**

— Klaus Schulten

Maryland, in the late 1980s, Bernard Brooks abandoned his effort, dubbed Gemstar, when Hewlett-Packard announced its blazingly fast 9000 series — which could be had for as little as \$12,000. Scientists are racing not just against each other, but against Silicon Valley.

Schulten has played that game before. In Munich in the late 1980s, he built his own parallel supercomputer out of 60 processors mail-ordered from England. He carried his computer in a backpack to his new laboratory in Illinois, where he ran a 30,000-atom simulation of the bacteriorhodopsin protein, which drives the photosynthetic reaction that turns light into an electric charge. His simulation lasted 263 picoseconds — less than a millionth of a millisecond — and required more than two years of continuous computation<sup>5</sup>. By then, his machine was obsolete.

### Thinking big

In the past 15 years, Schulten's ambitions have grown: from 100,000 atoms in 1999, to 300,000 in 2003, and culminating with his million-atom simulation of the tobacco mosaic virus published in 2006. To match his models, Schulten developed software that could scale with advances in parallel computers, something CHARMM could not do at the time. Chemist Richard Hilderbrandt, who supported the early development of NAMD at the computing directorate of the US National Science Foundation, says that the idea “was to take a large molecule and break it up into patches to distribute to processors. It was quite a bold step”.

The drawback of Schulten's strategy was that it could not simulate the behaviour of smaller molecules significantly faster than it could large ones. “If you have a protein of 500 atoms,” he says, “it's very difficult to put it on a parallel computer with 5,000 processors.”

Schulten emphasizes that his publicly funded group had to focus on ensuring that NAMD, which is freely distributed, would run on a wide range of platforms. Shaw's team, in contrast, could tune Desmond for its state-of-the-art computing cluster, about a year before similar clusters were available at National Science Foundation computing centres.

Shaw says that profits are a long way off, and that he is working to share his team's technology

## From science to finance and back again

For a man whom *Fortune* magazine once named King Quant, David Shaw does not come across as particularly regal. “I never understood why, if you want to be accepted in the business community, you have to wear something that restricts blood flow,” he says, tieless and with his top shirt button undone, in his office at the Columbia University Medical Center in New York.

Shaw's integration of science and trading may have been predestined: his stepfather was a professor of finance at the University of California, and his biological father was a plasma physicist who worked in the defence industry. For his part, the younger Shaw founded a technology consulting company while an undergraduate at Stanford University and, even after he joined Columbia University with a generous faculty package, he approached venture capitalists in the hope of getting \$10 million to develop his own parallel computing venture.

Instead of bringing money into his lab, Shaw got sucked into quantitative finance, where investors study the mathematical behaviour

of the markets to plan their strategies. In 1986, Morgan Stanley hired him to run its technology team, where he stayed for two years before founding his quantitative hedge fund, D. E. Shaw, one of the first to use sophisticated algorithms to exploit inefficiencies in the marketplace. “It was a field in its relative infancy and what that means to an academic type is there's low-hanging fruit,” he says. “This was part of the attraction: we could discover things no one else had found.”

His company became widely known for its 18% annual returns and a highly selective recruiting process in which only 1 in every 250 candidates got selected, many of them PhDs in the hard sciences or winners of prestigious maths competitions.

Even on Wall Street, Shaw never strayed too far from science. Always a major contributor to the Democratic party, in 1994 Bill Clinton appointed him to the President's Council of Advisors on Science and Technology, where he served for seven years and was charged with improving the use of technology in schools.

But as Shaw's 50th birthday neared in 2001, he began to look for an exit. His company had more than 1,000 employees, and Shaw was no longer engaged in the quantitative problem-solving that fascinated him. His sister was battling breast cancer — she died in 2003 — and Shaw believed he could contribute to medicine, not just financially but intellectually. In his spare time, he had been reading up on the computational puzzles of molecular dynamics and talking with academic friends.

In October 2002, Shaw hired his first computer scientist, trained at the Massachusetts Institute of Technology, to manage operations at D. E. Shaw Research. The venture, which Shaw compares to a tiny and highly focused Bell Labs, now has nearly 60 staff. Some researchers in the field are still getting used to the newcomer, but Shaw does not see science as a contest. “Maybe it was because I was in the financial field,” he says. “It's a zero sum game — you cannot make money unless someone else is losing money. That's one of the reasons why I like science: it doesn't work that way.”

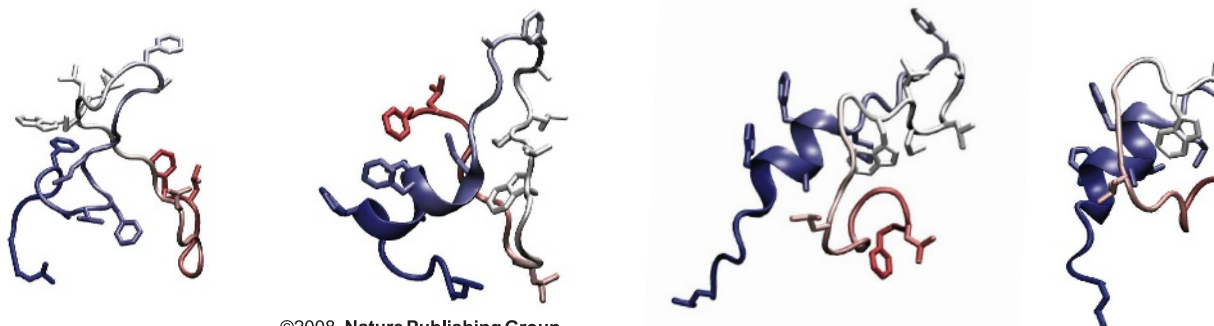
B.B.

as much as possible. But his proprietary algorithm will ultimately be sold to industry through an agreement with Schrödinger, a biotechnology company founded by chemist Richard Friesner, a colleague of Shaw's at Columbia. Schulten had only inklings of Shaw's ambitions when he gave a seminar at D. E. Shaw Research in October 2004. “At that time it was clear that there was a competition,” he says, “but in a very civilized way.” Even so, he

says, “I wouldn't have told them about a great solution I had developed, and they wouldn't tell me their solution.”

Although Schulten's software has been a boon to many researchers, with a development cost of \$20 million it might also be considered a drain on their resource pool. Some scientists contend that the pursuit of speed has hindered alternative modes of inquiry. “I think it's unfortunate that some of the researchers who use

Twist in the tale:  
a simulation of some steps  
in the folding of the villin  
headpiece, one of the  
fastest-folding proteins.



R. LEMOINE  
more established codes with a broader range of functionality are not getting the same access to national resources,” says computational chemist Charles Brooks of the Scripps Research Institute in La Jolla, California.

### Tough decisions

Some participants at a 2001 supercomputing conference recall Hilderbrandt telling the audience that users should switch from older programs such as CHARMM to modern parallelized packages, such as NAMD. Hilderbrandt, who is now at the Department of Energy, does not recall being so specific, but says he still believes NAMD is “the program of choice” for most applications.

Michael Crowley at the National Renewable Energy Laboratory in Golden, Colorado, doesn't buy that. He uses CHARMM to study biofuels and says: “CHARMM has functionality that as far as I know, no other program comes near.” He says that when he has applied for supercomputing time from allocating agencies, “you can almost expect that somebody is going to suggest you use NAMD”.

There are deeper questions about the pursuit of ever-longer timescales. “It's clear to me that what's emerging out of both Schulten's and Shaw's efforts are technological advances that are going to affect the entire community,” says Brooks. “But whether an individual achievement of a millisecond timescale for any particular simulation is of great significance, I'm not entirely sure.”

Vijay Pande, at Stanford University in California, has pioneered the folding@home distributed-computing project, which uses the personal computers and Sony PlayStations of more than 250,000 volunteers to study protein folding. “The revolution that's going on,” he says, “is people are now treating molecular dynamics in a much more sophisticated way, where they are running hundreds or thousands or millions of simulations and then data-mining those simulations.” Because a simulation may take a slightly different course each time, he notes, a single long simulation cannot provide the statistical information that

**“Researchers who use more established codes are not getting the same access to resources.”**

— Charles Brooks

must be gathered over many runs, such as the affinities for binding to a drug.

Schulten and Shaw may also be pushing current models to their breaking point. Neither group is investing significant resources in improving fixed-charge force fields, which might turn out not to be accurate enough for lengthy simulations. For instance, when two atoms approach one another, the electron orbits of one can get sucked towards the positive charge generated by the other. This phenomenon, called polarizability, is cumbersome to model and slow to compute. Shaw estimates that it would slow down computation by roughly a factor of ten; Schulten thinks it may be only a factor of two.

Yet these difficulties may be a reason for moving forward, not calling a halt. Longer simulations can show where the models are failing, and they can guide the distributed-computing approach. Shaw believes his group can make a meaningful contribution to the field, but he is well aware of the problems ahead. “If you have something you're sure is going to work,” he says, “you're not being ambitious enough.”

Last year, Schulten's group started running a new version of NAMD that can handle smaller

molecules faster. His team has also started programming the graphics accelerator chips prized by PC gamers — an economical solution to the hardware problem that could further shrink Anton's expected lead. And, now that the team is up to speed with the University of Illinois's cluster, Abe, it has tailored a special version of NAMD to compete on equal terms with Desmond.

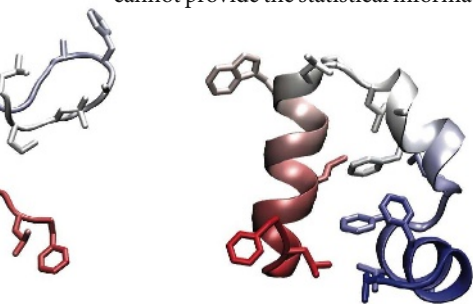
Two months ago, Schulten was delighted to tell Shaw about a simulation of a 38,000-atom protein, in which NAMD had set a new personal best, computing a 0.1-microsecond simulation in the course of a day. “We agreed, now the programs are pretty equal,” says Schulten. And for his part, Shaw may be starting to concede that each algorithm has its benefits. “Schulten has made extraordinary strides in his NAMD code,” he says, “so it's not obvious to me that Desmond will be significantly faster for all applications.”

**Brendan Borrell is a freelance science writer in New York City.**

1. Freddolino, P. L. *et al. Structure* **14**, 437–449 (2006).
2. Bowers, K. J. *et al. Proc. ACM/IEEE Conf. on Supercomputing (SC06)*, Tampa, Florida, 2006.
3. McCammon, J. A., Gelin, B. R. & Karplus, M. *Nature* **267**, 585–590 (1977).
4. Arkin, I. T. *et al. Science* **317**, 799–803 (2007).
5. Heller, H., Schaefer, M. & Schulten, K. *J. Phys. Chem.* **97**, 8343–8360 (1993).



**Number cruncher:** David Shaw has used his computer skills to make money and model proteins.



## Citations: rankings weigh against developing nations

SIR — Scientists and whole institutes are frequently judged by the number of citations of their papers in scientific journals, and project funding depends on it. But, as Clint Kelly and Michael Jennions note in Correspondence ('H-index: age and sex make it unreliable' *Nature* **449**, 403; 2007), the context and relevance of citations are crucial in reaching this judgement.

Researchers from developing nations often face another problem. In the name of local issues and the national interest, they are required to publish in national journals that rarely find a place among cited journals and have a very limited circulation abroad.

For example, a study of the Thomson Scientific Essential Science Indicators (ESI) during the past five years has found that the National Geophysical Research Institute (NGRI) in Hyderabad, India, scores among the top 1% of institutions publishing in the geosciences. During this period, the NGRI had 2,338 citations of 657 papers ([www.in-cites.com/institutions/2007menu.html](http://www.in-cites.com/institutions/2007menu.html)). But if it had not published more than half its publications in national journals — not all of which figure in the ESI database — the NGRI could have been ranked even nearer the top.

In formulating their criteria, publications from institutes and by individuals in local and national journals should also be taken into account: this could be done by assigning some weighted average. The total number of publications in national journals not counted by the ESI would then be considered and weighted in order to arrive at a more appropriate index.

**D. C. Mishra**

National Geophysical Research Institute, Uppal Road, Hyderabad 500 007, Andhra Pradesh, India

## Citations: poor practices by authors reduce their value

SIR — On 22 November, the Higher Education Funding Council for England announced that the assessment and funding of science-based disciplines will in future be "based on citation rates per paper, aggregated for each subject group at each institution" ([www.hefce.ac.uk/Pubs/HEFCE/2007/07\\_34/07\\_34.pdf](http://www.hefce.ac.uk/Pubs/HEFCE/2007/07_34/07_34.pdf)).

Changes in performance indicators always strongly influence individual and institutional behaviour and 'citation game-playing' will no doubt become a staple of coffee-room conversation. What is less clear is how the citation practices of authors may influence bibliometric indicators.

Citation practices are known to be imperfect. The documented problems

include excessive citation of an author's own work. Papers cited can be inappropriate or ambiguous in their support and, in some cases, the authors may not have read the papers they cite. Authors may form 'citation coalitions' within research networks. They may fail to provide citations to intellectual precursors or to work reporting conflicting conclusions. There are geographical and language biases. The increasing number of many-authored papers makes it impossible to have a clean-cut general metric in which one author is associated with one paper.

Taken together, these factors represent a problematic degree of error for the proposed bibliometric system of assessment. They place added responsibility on journal editors and reviewers as arbiters of appropriate author conduct.

Unfortunately, there are no simple solutions. Currently, identifying poor citation practices is not emphasized in the peer-review process, so perhaps journals could adopt a system of random citation audits, or periodically request evidence of citation appropriateness from authors. In reality, time constraints and the sheer volume of submissions to many journals mean that such measures are unlikely to be implemented soon.

Until referencing practices improve, we would argue that using citation rates to assess performance is fundamentally flawed.

**Peter A. Todd\*, Richard J. Ladle†**

\*Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, 117543, Singapore

†Oxford University Centre for the Environment, Dyson Perrins Building, South Parks Road, Oxford OX1 3QY, UK

## Glacier programme shows the value of 'ground truth'

SIR — On-the-ground monitoring is undervalued, as Euan Nisbet points out in his Commentary 'Cinderella science' (*Nature* **450**, 789–790; 2007). Long-term monitoring data provide the critical foundation we need in order to develop an understanding of the processes at work. This, in turn, enables modelling studies and rationally based management decisions. That is why having 'ground truth' — information gathered on the spot — to combine with satellite observations and modelling is even more critical today.

Through the 1980s and 1990s, we saw a deterioration in many key long-term monitoring programmes, the best example being the reduction in the number of US Geological Survey gauging stations. In recent years there has been a push to increase such networks as the US Natural Resources Conservation Service's snowpack telemetry gauging stations. We can see how hard it is to construct long-term

records on sea-level rise, because tide gauge records have seldom been continuous.

Monitoring is as important, I believe, as expanding the horizons of research. The data sets gained are key to the expansion of knowledge as well. I monitor the mass balance of more glaciers than any other programme in North America and have done so for 25 years without any federal money. This was crucial from the start — I was correctly informed that the federal government was not interested in funding long-term monitoring. As a result, I sought alternative funds that were sustainable but also enforced the use of cost-efficient techniques. Both have been key to maintaining the extensive annual fieldwork programme that is required to measure and report glacier mass balance.

**Mauri Pelto**

North Cascade Glacier Climate Project, Nichols College, Dudley, Massachusetts 01571, USA

## Restricted access to fossils hinders claim confirmation

SIR — Your Editorial 'Replicator review' (*Nature* **450**, 457–458; 2007), detailing the logic needed to evaluate reports of major research breakthroughs, such as the recent paper on the transfer of nuclear material in a primate, is commendable. It is responsible to require independent confirmation of 'extraordinary claims', in particular for those that are difficult to reproduce.

However, unique materials, such as fossils, require scrutiny by independent researchers to evaluate similarly extraordinary claims. Gaining access to these can be highly problematic. This issue is particularly pervasive in palaeoanthropology, where newly described fossil materials are often barred from review after initial reports. Your News story 'Anthropologists rocked by fossil access row' (*Nature* **428**, 881; 2004) gives one example. Given that *Nature* is the preferred outlet for analysis of palaeontological discoveries, the editors are in a position to encourage broader access to these valuable specimens.

**Christopher P. Heesy**

Department of Anatomy, Midwestern University, 19555 North 59th Avenue, Glendale, Arizona 85308, USA

Readers are welcome to comment at [http://blogs.nature.com/peer-to-peer/2007/11/peerreview\\_for\\_strong\\_claims\\_1.html](http://blogs.nature.com/peer-to-peer/2007/11/peerreview_for_strong_claims_1.html)

**Contributions to this page may be submitted to [correspondence@nature.com](mailto:correspondence@nature.com). Published contributions are edited. We welcome comments on publishing issues at Nautilus (<http://blogs.nature.com/nautilus>).**

## BOOKS &amp; ARTS

# Twenty-first-century anatomy lesson

Polymath pieces together the surprising past of the human body from fins, wings, hangovers and hiccups.

## Your Inner Fish: A Journey into the 3.5-Billion-Year History of the Human Body

by Neil Shubin

Allen Lane/Pantheon: 2008. 240pp.  
£20/\$24

### Carl Zimmer

Six hundred years ago, anatomists were rock stars. Their lessons filled open-air amphitheatres, where the curious public rubbed shoulders with medical students. While a surgeon sliced open a cadaver, the anatomist, seated above on a lofty chair, deciphered the exposed mysteries of the bones, muscles and organs.

Modern anatomists have retreated from the stage to windowless medical-school labs. They have ceded their public role to geneticists unveiling secrets encrypted in our DNA. Yet anatomists may be poised for a comeback, judging from *Your Inner Fish*. Neil Shubin, a biologist and palaeontologist at the University of Chicago, Illinois, delves into human gristle, interpreting the scars of billions of years of evolution that we carry inside our bodies.

I met Shubin ten years ago while writing a book about major transitions in evolution. At first glance, his lab suggested a person who had yet to make up his mind just what kind of scientist he was going to be when he grew up. Shubin spent much of his time studying fossils of mammals and other creatures he dug up in places such as Canada's Bay of Fundy. He also stained embryos to learn about the mysterious process by which limbs develop into fins, legs, wings and hands.

Actually, Shubin's mix of research had a focus: he wanted to understand how new structures evolve. How, for example, could the tetrapod limb arise from lobe-fin fish that had no trace of hands or feet? Shubin combined information from both fields to identify the genes that changed during these key evolutionary transitions. In the late 1990s, this 'integrative biology' was radical. It ran counter to the long tradition of specialization in the field. Other developmental biologists who had spent decades poring over shark embryos did not think of heading off to the mountains to find fossils to study.



Author Neil Shubin (above) discovered the transitional fossil *Tiktaalik roseae* (below).



A decade later, Shubin has plenty of company. Journals regularly publish reports on the synthesis of fossils, genes and embryos. Fossils of whales with legs have helped scientists figure out which genes changed as whale legs gradually disappeared. Tinkering with bat embryos has suggested how their hands stretched into wings. Shubin's own work on limbs has moved forward spectacularly. In 2006, he and his colleagues made international headlines with the discovery of the transitional fossil *Tiktaalik roseae*. This 370-million-year-old fish had acquired most of the tetrapod limb in its stout

fins, including some wrist bones. And while Shubin and his colleagues were digging up *Tiktaalik* in the Arctic, some of his students stayed behind in Chicago to find equally useful clues about the transition from sea to land in the genes that help build the fins of sharks and paddlefish.

*Your Inner Fish* combines Shubin's and others' discoveries to present a twenty-first-century anatomy lesson. The simple, passionate writing may turn more than a few high-school students into aspiring biologists. And it covers a lot of ground. Shubin inspects our eyeballs, noses and hands to demonstrate how much we have in common with other animals. He notes how networks of genes for simple traits can expand and diversify until they build new complex structures such as heads. Also, that hangovers explain how our ears evolved from sensory cells on the surface of fish. He investigates the hic-

cup, the result of a tortuous nervous system.

Some of the case studies will be familiar to those who have read a lot about evolution, but most readers will find some surprises. I learned that in sharks, the testes sit near the head. As male human embryos develop in the womb,

their testes gradually descend from that ancestral position to wind up in the scrotum. As they migrate, they push down on the body wall, creating a weak spot. It is here that the intestines can slip through during a hernia.

Along the way, Shubin offers some striking examples of how science works. He did not wander in the Arctic hoping to trip over a fossil of a transitional species. He knew from previous discoveries exactly which formations he should look for — mid-Devonian sedimentary deposits. When his colleagues began to unearth *Tiktaalik*, a glance at its distinctively flat skull confirmed that they had found what they had come for. They had learned their anatomy well.

Carl Zimmer is a science writer based in Guilford, Connecticut, and is author of *Microcosm: E. coli and the New Science of Life*.

T. DAESCHLER, ACADEMY OF NATURAL SCIENCES OF PHILADELPHIA/J. WEINSTEIN, FIELD MUSEUM

# Interdisciplinary inspiration

## Artsience: Creativity in the Post-Google Generation

by David Edwards

Harvard University Press: 2008. 208pp.  
£12.95, \$19.95

### Alice W. Flaherty

What if science could move us in the same way art does? What if art could have the social impact of a technological advance? David Edwards' slender book proposes that they can. A professor of biomedical engineering at Harvard University, Edwards has launched several programmes with creative and humanitarian missions. *Artsience* is in some respects a part of his most recent project, Le Laboratoire (see *Nature* **449**, 789; 2007). This Parisian cultural centre, which opened last October, recruits scientists and artists to interact, to spur their innovation and to engage the public in the process.

The 'post-Google' subtitle presents the book as part of a wave of technological progress — Edwards coined the term 'artsience' as if to suggest the emergence of a new discipline. But the book is less a technical tool than a motivational one: an exhortation for interdisciplinary intellectuals.

Most of the book sketches vivid models of men and women who passionately mix art and science. They include a pianist whose PhD in electrical engineering spurs her to compose music using chaos theory, an infectious-disease researcher who mingles theatre about Chekhov's tuberculosis with public-health advocacy, and a mathematician whose visual imagery drives both his paintings and his fluid-mixing models.

The profiles, apart from their notable freedom from gender bias, take the Great Man approach to understanding creativity that has been championed by researchers such as Howard Gardner and Dean Simonton. That said, Edwards's book is surprisingly ahistorical. Where are the hoary 'artsience' greats such as Leonardo or Goethe? Instead, Edwards's life-sketches encourage us that polymathic creative lives are possible even in today's era of subspecialization.

His contemporary artsientists have lives the reader might emulate — something that very dead Renaissance men do not. Edwards infects us with his subjects' creativity. When the final



These works by artist Fabrice Hyber were inspired by a visit to polymer scientist Robert Langer's lab.

chapter turns from vignettes to his utopian Laboratoire, we're rooting for it to succeed.

Le Laboratoire aims to foster the quality of the creative process, and de-emphasizes pressure for results. Practising scientists, whose process-to-product ratio is inevitably high, might favour such an emphasis. Yet many scientists — and all grant agencies — feel otherwise. A product-oriented reader might point out that without the constraint of a need for results, most novel attempts at 'artsience', however fervent, could end as badly as the works on view at the Museum of Bad Art ([www.museumofbadart.org](http://www.museumofbadart.org)).

Edwards might reply that the creation of what's new and good inevitably generates a great deal of what's new and bad, just as sex produces more failed offspring than vegetative replication does. Intrinsic pleasures of 'process', such as curiosity, turn out to drive creative results more strongly than extrinsic rewards do.

Creativity researcher Teresa Amabile and her colleagues at Harvard have shown that even positive results such as praise and being paid can decrease inventiveness, by distracting the creator from the process of creation.

Does the creative process differ in science and art? Edwards examines the traditional dichotomy between the artistic method (associative, emotional, and vividly image-based or sensual) and the scientific method (deductive, rational and symbolic). It is not surprising when his case studies knock that straw man down. Researchers in creativity would call those poles 'primary process' and 'secondary process' thought. Each is important for creativity in both science and art. In either domain, primary process produces a novel idea, and secondary process refines or edits it.

Edwards also examines the traditional split between art as pleasing but impractical and science as useful but arcane. In most of his 'artsience' examples, art works to make science more accessible, whether to the scientists themselves, to entrepreneurs who might translate ideas into reality, or, ultimately, to the public. Indeed, some reviewers have interpreted Le Laboratoire as a concept-heavy science museum.

Many scientists who find popularized science distastefully sloppy need not worry that Le Lab will attract the hoi polloi. Its website

([www.laboratoire.org](http://www.laboratoire.org)), whose avant garde videos echo those from Andy Warhol's Factory, does not seem aimed at mass consumption. It does promise a programme that can kindle scientists and artists to burn

more brightly, and may inspire new ideas in a way that a more specialized centre would not. Reading this book, for all its slenderness of content, may do the same.

Alice W. Flaherty is assistant professor in the Department of Neurology at Harvard Medical School, and Director of the Movement Disorders Fellowship at Massachusetts General Hospital, Boston, Massachusetts, 02114, USA. She is the author of *The Midnight Disease: The Drive to Write, Writer's Block, and the Creative Brain*.

M. DOMAGE/D. FAUST

# Biography of a blockbuster text book

## The Anatomist: A True Story of Gray's Anatomy

by Bill Hayes

Ballantine Books: 2007. 272pp. \$24.95

### Ken Arnold

We've all heard of it. Many of us have flicked through it in a bargain book shop. It has gone through more than 30 revised editions on each side of the Atlantic and has sold more than five million copies. *Gray's Anatomy* is surely one of the world's great books. But, as Bill Hayes discovered in researching this publishing marvel, evidence of how it came about is scant.

Illustrated anatomy texts had been in circulation for more than half a millennium when *Gray's Anatomy* was published in 1858. Its author, English surgeon Henry Gray, aimed not to produce an enduring classic, but to improve on the passable text books he had used as a student at St George's Hospital Medical School in London.

The medical curriculum's recent expansion and the increasingly widespread use of anaesthesia provided a fertile context in which to launch a fresh anatomical text book. Arguably Gray's most significant innovation was to focus on surgical anatomy, ensuring that his book would remain useful to medics long after they had entered the professional world. This commercial formula has proved buoyant ever since.

From its first reviews, critics were struck by the clarity and functionality of the atlas's pictures. Such fare had long served to objectively analyse the body in ever finer detail and to remind scientists, doctors and patients alike of its subjective and emotional resonance. Gradually the rigorous demands of the former squeezed out the opportunity to indulge in the latter.

*Gray's Anatomy* effectively marked the end of the road for the troops of playful cadavers that had, in earlier volumes, cavorted with props and danced as only the dead know how. Here instead images shied away from the notion of style altogether. The book offered a set of pictures that students and professionals were supposed to look through rather than at, into the realities of nature that they revealed. Recently, medical thinkers have begun to ponder what was lost when the two approaches were separated, and whether a third way — medical humanities — should now be cultivated.

Gray's bible of medical understanding emerged from his collaboration with another Henry. In June 1850, Gray, the project's instigator, invited the more junior Henry

Vandyke Carter to supply what became the iconic illustrations. Even before their inspiring collaboration, Carter prophetically declared: "Two persons are generally concerned in every fact, one discovers part, the other completes and corrects."

Of Gray, we know very little — even the year of his birth is contested. Luckily, various archives reveal much more of Carter's life and work. The illustrator probably inherited his aesthetic abilities from his father, the practising Scarborough artist Henry Barlow. Carter headed south to pursue medical studies in London and took to anatomy with a passion, spending whole days dissecting. The combination of his skills as a draftsman and the depth of his anatomical knowledge recommended him to Gray.

The inspired collaboration lasted for just one project. By the time the work was published, Carter had moved to Bombay; here he clocked

up an extended spell in research and administration at the Grant Medical School, ending up as its principal. Hayes is as concerned with character as career, and his lively prose provides much insight into Carter's colourful but failed romantic entanglements. But we never really get much insight into just what made Carter's drawings so compellingly distinctive.

*The Anatomist* also concerns the progress of a third anatomist: Hayes himself. Early on in his research, Hayes was determined that he too should learn through scalpel and cadaver as well as lecture, library and archive. Some of his most mem-

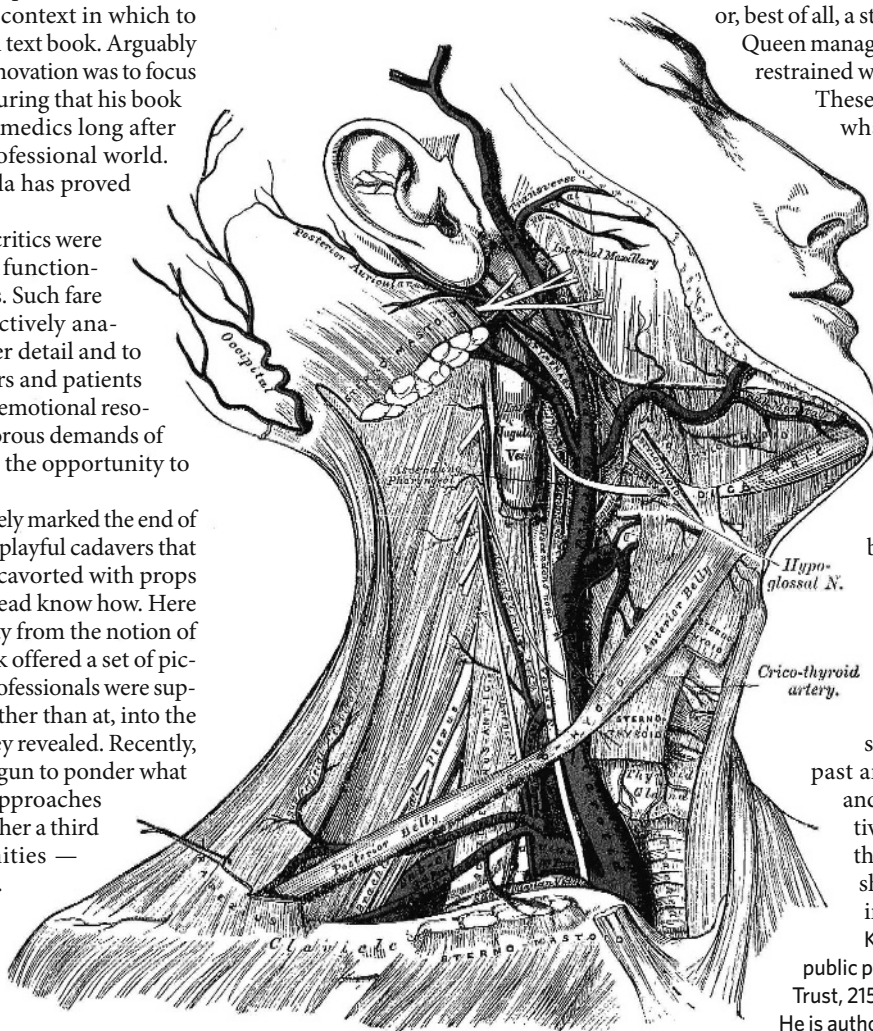
orable writing describes the dissection classes he attended in San Francisco. We are treated to a selection of fascinating anatomical snippets about, for example, how to trace evidence of the sealed hole in the fetal heart through which the mother's blood enters; or how to find the kidney in a cadaver; or that blood flowing out of the heart is first used to feed the heart itself; or, best of all, a structural analysis of how the Queen manages to deliver such a uniquely restrained wave.

These sections allow Hayes to do what seemingly every writer must these days: he tells us about himself. Those tempted to skip over these fashionable journalistic passages might actually profit from lingering over them. It is here that Hayes really comes to grips with the emotional tension inherent in anatomical studies: the way in which layers of a dead body can be stripped away so we might better understand life.

An important work of medical history *The Anatomist* is not. It is, though, an enjoyable contribution to the burgeoning field of medical humanities, skillfully bringing together past and present, objective facts and speculations, in a provocative meditation on a text book that might well still be helping shape young medical minds in another 150 years. ■

Ken Arnold is head of public programmes at the Wellcome Trust, 215 Euston Rd, London NW1 2BE. He is author of *Cabinets for the Curious: Looking Back at Early English Museums*.

**"The emotional tension in anatomy: layers of a dead body stripped to better understand life."**



Neck arteries, illustrated by Henry Vandyke Carter, from the 1858 edition of *Gray's Anatomy*.

## BEHAVIOURAL NEUROSCIENCE

# Neurons of imitation

Ofer Tchernichovski and Josh Wallman

**In songbirds, a class of neurons shows a striking similarity in activity when the bird sings and when it hears a similar song. This mirroring neuronal activity could contribute to imitation.**

Songbirds are champion mimics. A nightingale, for example, can imitate at least 60 different songs after a few exposures to each<sup>1</sup>. A young bird learns its species' song through imitation, and the ability is also socially important: a bird on its territory will often respond to an intruder's song by singing a similar song, thus acknowledging the intrusion<sup>2</sup>. What neurons might mediate these imitative and communicative powers? On page 305 of this issue, Prather *et al.*<sup>3</sup> identify a class of brain neurons that are active both when the bird hears a song and when it replies by singing a similar song.

As such, these neurons are reminiscent of the mirror neurons discovered in the monkey brain. These respond similarly whether an action is perceived or performed, and they aroused enormous interest as a possible key to understanding such disparate phenomena as imitation and empathy. Mirror neurons are activated both when a monkey performs a discrete action — such as grasping a small object between thumb and forefinger — and when it sees another monkey or a human do the same<sup>4</sup>, but not when the same

action is performed without accomplishing the goal (pretending to grasp the object).

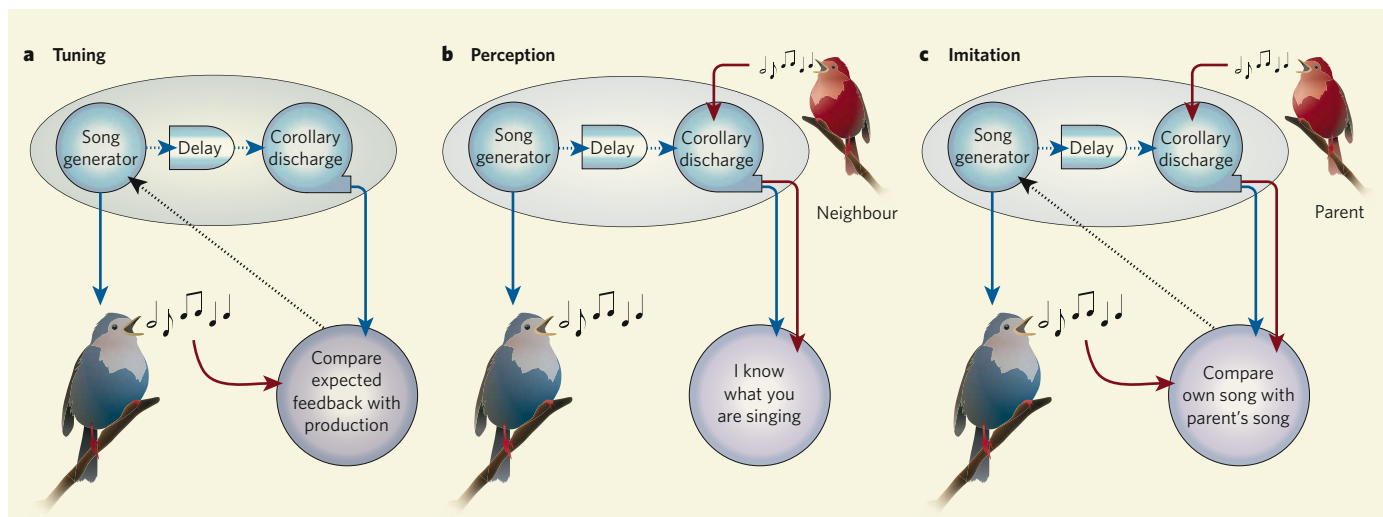
To mirror neurons, actions performed or observed are equivalent, so they could mediate imitation — a most mysterious form of learning. How does one know what pattern of muscle contraction corresponds to a particular visual effect? The psychologist William James speculated that infants correlate their random limb movements with the sight of their limbs, thereby forming an association between motor outputs and visual inputs that allows them to infer how others make similar limb movements. But one does not need to spend hours in front of a mirror to imitate the facial expressions of others<sup>5</sup>; nor do French or Italian children need to observe themselves to acquire the facial gestures characteristic of their elders. Mirror neurons may be the link between the sensory information perceived and gestures produced.

Mirror neurons might also facilitate our perception and memory of complex sensory stimuli<sup>6</sup>. For example, a sequence of familiar dance steps could be more easily encoded

in memory in terms of the commands that the brain sends to move the limbs than it could by remembering all the small visual changes these limb movements produce. This function of mirror neurons would not be independent of their ability to facilitate imitation. Indeed, it is a common experience, when watching a car chase in a film, to feel oneself involuntarily making small steering or braking movements.

The responses of mirror neurons have led psychologists to propose that they provide a way of inferring the workings of another's mind, and so are essential for the development of social communication and empathy<sup>7</sup>. This has put the emphasis on mirror neurons' higher-level functions. The mirroring neurons Prather and colleagues found in songbirds may also have such functions, but they seem to have more prosaic roles in acquiring motor skills and in learning.

All the likely functions of the songbird's mirroring neurons are related to singing. The neurons are located in the brain's principal song-generating nucleus, the high vocal



**Figure 1 | A singing-listening neuronal connection.** The neurons identified by Prather and colleagues<sup>3</sup> could be involved in three sensorimotor processes. **a**, The delayed corollary discharge of song patterns can be simultaneously compared with auditory feedback of the bird's own song, allowing tuning. **b**, The auditory responses (in the mirroring neurons) to songs of a neighbour might be compared with the memory of the corollary

discharge produced during singing. This might allow the bird to identify an imitation by that neighbour. **c**, Corollary discharges while singing might be compared with a memory of the mirroring neurons' response to the parent's song. The error may then feed back to the song generator and guide vocal learning during song development, in addition to guidance from auditory input during singing (lowest arrow).

centre (HVC). Like other neurons in the HVC, they respond to specific songs with highly stereotyped timing of nerve impulses. Curiously, when the bird is singing, these mirroring neurons are deaf to auditory input, meaning that their responses switch between being auditory and being a reflection of motor activity.

Because the HVC is a premotor structure, it would be expected that nerve impulses would occur here earlier than the resulting sounds, whereas the auditory responses of the neurons would occur later. But Prather *et al.*<sup>3</sup> find that the timing of nerve impulses from the mirroring neurons of the HVC is the same whether the bird is singing or listening. This remarkable delaying of the motor signal implies that the mirroring neurons are providing a 'corollary discharge' signal, that is, a neural representation of the motor output (the song being sung) encoded in a way that can be readily compared with the auditory input (hearing the song). Thus, these neurons present two solutions to the brain's main problems in comparing motor outflow with sensory inflow: they form an equivalence between the motor output and the resulting sensory feedback, and they compensate for the delay between them<sup>8</sup>.

What functions might this corollary discharge have? Prather and colleagues found a clue by investigating where the projections (axons) of the mirroring neurons go. The HVC has two outputs: one down the motor song pathway to the vocal organ, and the other to the anterior forebrain pathway (AFP), which is required for song learning but not for singing. All of the mirroring neurons project to the AFP, which, in turn, trains the motor song system during song learning by introducing variability into the song patterns<sup>9</sup>.

Sending corollary discharge into the AFP might have several functions. First, synchronous responses to hearing and singing might allow tuning of the song (Fig. 1a). While singing, the corollary discharge from the song generator might be compared with the auditory feedback from the resulting song. Such an online comparison might allow adjustments of the song produced<sup>10</sup>. Second, when a bird hears a neighbour imitating its song, its mirroring neurons might send a pattern to the AFP similar to that of the corollary discharge (Fig. 1b). The AFP might then recognize the song, thereby providing an efficient mechanism for the bird to identify its neighbour.

Third, mirroring neurons could be necessary for the gradual process of the bird learning to imitate the songs of its parent (Fig. 1c). The young bird might compare the corollary discharge of its singing with the memory of the responses of the mirroring neurons to the parent's songs, thereby simplifying the comparison and facilitating a gradual improvement in the imitation. Possibly related to this function is that, during the several weeks that song learning takes, many HVC neurons are replaced by others<sup>11</sup>. The mirroring neurons identified by Prather *et al.*

belong to a population that is not replaced, but is stable across song development. It is tempting to imagine that this stability keeps the corollary-discharge signal reliable while the song produced is changing, thereby defining a role for these neurons at the centre of the sensorimotor convergence that facilitates vocal imitation<sup>8</sup>.

The exciting findings of Prather *et al.*<sup>3</sup> offer the possibility of following the emergence of sensorimotor mirroring as the song becomes increasingly structured and similar to the song being learned. More generally, the mystery of how a neuron can have similar responses to performing and experiencing an action might be clarified by studying which response develops first and how the two responses converge, resulting in a common neural representation. ■

Ofer Tchernichovski and Josh Wallman are in the Department of Biology, The City College of

New York, 138th Street and Convent Avenue, New York, New York 10031, USA.  
e-mail: ofer@sci.ccny.cuny.edu

1. Hultsch, H. & Todt, D. *J. Comp. Phys. A* **165**, 197–203 (1989).
2. Beecher, M. D., Campbell, S. E., Burt, J. M., Hill, C. E. & Nordby, J. C. *Anim. Behav.* **59**, 21–27 (2000).
3. Prather, J. F., Peters, S., Nowicki, S. & Mooney, R. *Nature* **451**, 305–310 (2008).
4. Rizzolatti, G., Fadiga, L., Gallese, L. & Fogassi, L. *Brain Res. Cogn. Brain Res.* **3**, 131–141 (1996).
5. Meltzoff, A. N. & Prinz, W. *The Imitative Mind: Development, Evolution, and Brain Bases* (Cambridge Univ. Press, 2002).
6. Craighero, L., Metta, G., Sandini, G. & Fadiga, L. *Prog. Brain Res.* **164**, 39–59 (2007).
7. Gazzola, V., Aziz-Zadeh, L. & Keysers, C. *Curr. Biol.* **16**, 1824–1829 (2006).
8. Troyer, T. W. & Doupe, A. J. *J. Neurophysiol.* **84**, 1224–1239 (2000).
9. Olveczky, B., Andalman, A. S. & Fee, M. S. *PLoS Biol.* **3**, e153 (2005).
10. Turner, E. C. & Brainard, M. S. *Nature* **450**, 1240–1244 (2007).
11. Scharff, C., Kirm, J. R., Grossman, M., Macklis, J. D. & Nottebohm, F. *Neuron* **25**, 481–492 (2000).

## INORGANIC CHEMISTRY

# Uranium gets a reaction

James M. Boncella

**The most common form of uranium in solution is notoriously unreactive, limiting the use of the element. But interactions of this complex with potassium ions unleash a potentially rich seam of unexpected chemistry.**

It's not often nowadays that a new chemical reaction is discovered, so Arnold and colleagues' report<sup>1</sup> (page 315) of some unprecedented uranium chemistry is a cause for celebration. They describe a reaction of uranyl ions ( $\text{UO}_2^{2+}$ ), the most common form of uranium in solution. Until now, almost all known reactions of these ions involved only the binding of molecules called ligands to the uranium atom, but Arnold *et al.* have found a way to force the oxygen atoms to react. This represents a sea-change in uranium chemistry, and could enable completely new methods to be developed for manipulating uranium compounds in solution.

Uranyl ions were discovered shortly after uranium itself. Their reactions are crucial for the extraction of uranium ore, the processing of nuclear fuel and the disposition and movement of uranium in the environment. The ions are characterized by the extreme thermodynamic stability of their uranium–oxygen double bonds ( $\text{U=O}$ ), which are very unreactive. As a result, almost all the chemistry of uranyl ions has been limited to changing the ligands that bind to the metal, leaving the  $\text{U=O}$  bonds unaltered. Although these ligand-exchange reactions are undoubtedly useful — they form the foundation of uranium processing — much effort has been directed towards finding ways to make the oxygen atoms react, mostly without success.

In many ways, the low reactivity of the uranyl ion is surprising. It is a member of a much larger class of complexes that includes reactive ions such as chromate and permanganate, which have oxygen atoms that readily form bonds to other molecules. These reactive transition-metal ions are commonly used as oxidizing agents in synthetic organic chemistry. By contrast, the uranyl ion does not readily oxidize organic substrates. Furthermore, the molecular structures of transition-metal oxides are very different from that of the uranyl ion — in transition-metal oxides, the angles formed between adjacent metal–oxygen bonds are acute, but the equivalent bond angle in the uranyl ion is 180°.

The structure and stability of the uranyl ion results from a unique confluence of electronic effects that lead to the formation of strong, unreactive  $\text{U=O}$  bonds<sup>2</sup>. Because uranium has a high atomic number, relativistic quantum effects influence the energies of electrons in its atoms. This causes 'non-valence' electrons (known in uranium as 6*p* electrons) that are normally found close to the atomic nucleus to reside in a relatively high-energy orbital with a large radius. The 6*p* electrons can therefore interact with a high-energy 'valence' orbital (the 5*f* orbital), generating a set of hybrid orbitals. The 5*f* orbital would not normally interact strongly with ligands bound to the metal, but the hybrid orbitals can form a strong, linear

bonding interaction with two small atoms such as nitrogen<sup>3</sup> or oxygen — exactly as seen in the uranyl ion.

Uranyl ions are an extremely rare example of compounds in which the presence of non-valence, core electrons dictates the observed structure of a molecule. Such linear bonding interactions are not possible in transition-metal oxides because they do not possess valence *f* orbitals, and because the relativistic effects in transition metals aren't large enough to force any of their core orbitals to participate in bonding in a similar way. The key to Arnold and colleagues' discovery<sup>1</sup> is that they have found a way to disrupt the bonding interactions that stabilize U=O bonds so that the uranyl group can take part in an atypical reaction.

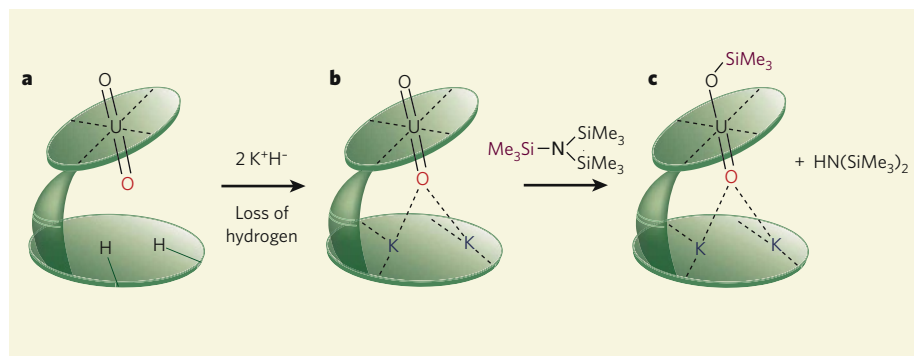
Arnold *et al.*<sup>1</sup> use a flexible ligand to simultaneously bind a uranyl ion and two potassium ions. The Pac-Man-like structure adopted by the ligand (Fig. 1) forces one of the uranyl oxygen atoms to donate electrons to the potassium ions. Under normal conditions<sup>4</sup>, interactions of this sort are not favourable; only the presence of the ligand scaffold makes it possible. The interaction with the potassium ions disrupts the characteristically strong U=O bonds, so that the uranium ion behaves as a strong oxidant<sup>5</sup>. The unbound oxygen atom is thus able to abstract a silicon-containing group from an organic substrate; the uranium atom accepts electrons (is reduced). Given the thermodynamic stability of the resulting silicon-oxygen bond, it is likely that the formation of such bonds also provides a substantial driving force for the observed reaction.

It is well documented<sup>6</sup> that the reactivity of transition-metal-oxide complexes can be tuned by changing ligands bound to the metal on the opposite side of the complex to an oxygen atom. Arnold and colleagues' process<sup>1</sup> is similar: the interaction of one of the oxygen atoms with metal ions affects the reactivity of the opposing oxygen atom. In this sense, the structure of the authors' ligand-ion

complex is similar to the active site of the enzyme cytochrome *c* oxidase, which converts oxygen molecules into water. Oxygen binds between two metal ions (one iron and one copper) in the enzyme active site, disrupting the bonding in the oxygen molecule and so facilitating a reduction reaction that cleaves the oxygen into two water molecules<sup>7</sup>.

You might think that any metal ion bound in the uranyl-ligand complex would be able to trigger Arnold and colleagues' reaction<sup>1</sup>, but this is not the case. Previous work<sup>8</sup> from the same group showed that several dipositive transition-metal ions — iron, manganese or cobalt — form bonding interactions to a uranyl oxygen atom when in complex with an appropriate ligand scaffold, but these complexes do not undergo the redox reaction observed for the potassium complex. This is puzzling, because potassium ions are not redox active, whereas the transition-metal ions are. Perhaps the greater size of potassium ions perturbs the U=O bond more than the smaller, transition-metal ions, thereby inducing the observed reaction. A crystal structure of the potassium-bound complex would help to clarify this, but it seems that the complex is not particularly stable, so its structure has not yet been determined. Further details of the reactivity and structures of the key compounds involved in Arnold and colleagues' reaction will undoubtedly be discovered as this intriguing chemistry is explored.

The authors' reported reaction<sup>1</sup> could perhaps be used as a strategy to manipulate uranyl ions in solution, but many questions must first be answered before its full potential can be realized. For example, can substrates other than silicon-containing compounds undergo reaction? Given that the current reaction is performed in an organic solvent, could this, or related chemistry, work in water, as would be needed for nuclear-fuel processing? And can this unusual reactivity be reproduced in oxide complexes of other, heavier actinide metals?



**Figure 1 | Uranyl-ion reaction induced by metal ions.** The oxygen atoms in uranyl ions ( $\text{UO}_2^{2+}$ ) are generally unreactive. **a**, Arnold *et al.*<sup>1</sup> show that uranyl ions bind to a rigid molecular scaffold (a ligand, shown schematically in green; dotted lines represent non-covalent binding interactions). **b**, The authors displace two hydrogen atoms in the uranyl-ligand complex with potassium (K, dark blue) ions. The potassium ions bind to the ligand close to one of the uranyl oxygen atoms (red), and so are forced to interact electronically with it. **c**, This interaction increases the reactivity of the remaining oxygen atom, which can remove a silicon-containing group ( $\text{SiMe}_3$ , purple, where Me represents a methyl group) from a substrate introduced into the reaction mixture. The uranium atom is simultaneously reduced.



## 50 YEARS AGO

"Team-work and discovery in science" — ... Dr. W. S. Kroll took issue with those who claim that the days of 'sealing wax-baling wire' science are over. The fight of the individual against the collectivity in which he lives is as old as humanity and it will never cease to exist. While Dr. Kroll granted that the team could not be avoided in development work, he challenged its justification in research ... He maintained that many laboratories in the United States are over-fond of gadgets and complicated equipment which often take more time to repair than to use. These instruments remove the investigator from his experiment ... We have to offer the recalcitrant lone-wolf research worker some asylum since he is now menaced with extinction. From *Nature* 18 January 1958.

## 100 YEARS AGO

"Public clocks and time distribution" — The interesting correspondence on "Lying Clocks" inaugurated by Sir John Cockburn in the *Times* has tended to degenerate into a display of advertisements by different firms interested in various systems of clock synchronisation ... [The] essential preliminary of the distribution of correct time signals is provided for by the Post Office authorities, working in cooperation with the Royal Observatory, Greenwich. The telegraphic service throughout the country is suspended for a few seconds, while the signal is sent through the trunk lines at 10 a.m. But, unfortunately, it is to be feared that the duty of forwarding this signal to the smaller towns is very carelessly and inefficiently performed ... If it were thoroughly well known that there did exist in every town and village an office where correct time could be had, even at some personal inconvenience, careful people would take the trouble to keep their clocks fairly accurate, and by so doing gradually educate the more indifferent to a higher standard.

From *Nature* 16 January 1908.

50 & 100 YEARS AGO

With the current interest in nuclear energy as a carbon-dioxide-free means of power generation, there is much interest in developing new methods for the chemical processing and reprocessing of uranium and other radioactive actinide elements. So the big question is whether Arnold and colleagues' discovery can be exploited in nuclear-fuel cycles of the future. It is much too early to say, but the fundamental chemistry that will be uncovered as we try to find out will be fascinating. ■

James M. Boncella is at the Los Alamos National Laboratory, Materials Applications and Physics Division, P.O. Box 1663, MS J514, Los Alamos,

New Mexico 87545, USA.

e-mail: boncella@lanl.gov

1. Arnold, P. L., Patel, D., Wilson, C. & Love, J. B. *Nature* **451**, 315–317 (2008).
2. Denning, R. G. *J. Phys. Chem. A* **111**, 4125–4143 (2007).
3. Hayton, T. W. *et al. Science* **310**, 1941–1943 (2005).
4. Sarsfield, M. J. & Helliwell, M. J. *Am. Chem. Soc.* **126**, 1036–1037 (2004).
5. Wilkinson, G., Gillard, R. D. & McCleverty, J. A. *Comprehensive Coordination Chemistry* Vol. 3 (Pergamon, Oxford, 1987).
6. Nam, W. *Acc. Chem. Res.* **40**, 522–531 (2007).
7. Qin, L., Hiser, C., Mulichak, A., Garavito, R. M. & Ferguson-Miller, S. *Proc. Natl Acad. Sci. USA* **103**, 16117–16122 (2006).
8. Arnold, P. L., Patel, D., Blake, A. J., Wilson, C. & Love, J. B. *J. Am. Chem. Soc.* **128**, 9610–9611 (2006).

## CANCER

# Hay in a haystack

Kevin M. Shannon and Michelle M. Le Beau

**Although some diseases occur when both copies of a gene are mutated, mutation of just one copy of certain tumour-suppressor genes promotes tumorigenesis. Identifying such mutations is arduous, but worth the effort.**

The myelodysplastic syndromes are thought to result from mutations in haematopoietic stem cells that result in the inefficient production of blood cells. Anaemia is a frequent manifestation, and patients often become dependent on red-blood-cell transfusions. These syndromes were previously called preleukaemia, because many affected patients ultimately progress to acute myeloid leukaemia. A subtype of the myelodysplastic syndromes, known as the 5q<sup>−</sup> syndrome, is characterized by loss of the q31–33 segment of the long arm of chromosome 5 (ref. 1), although the specific gene (or genes) within this region that are responsible for the disease are unknown. On page 335 of this issue, Ebert *et al.*<sup>2</sup> now pinpoint a culprit gene in the 5q<sup>−</sup> syndrome.

A cornerstone of modern cancer biology is Knudson's two-hit hypothesis<sup>3</sup>, which postulates that the inactivation of both copies (alleles) of a tumour-suppressor gene has an essential role in cancer development. Indeed, this 'biallelic' inactivation of tumour-suppressor genes such as *RB1*, *TP53*, *APC*, *BRCA1*, *PTEN* and *NF1* is fundamental to tumorigenesis.

Uncovering further tumour-suppressor genes is a major priority for understanding cancer biology and developing new therapies. This process typically begins with identifying a discrete DNA segment that is likely to harbour a tumour-suppressor gene. Techniques used include performing linkage studies in familial syndromes that predispose patients to cancer, identifying the boundaries of recurring cancer-associated deletions, and using markers to define domains in which tumour cells show absence of one germline allele (also known as loss of constitutional heterozygosity). By integrating data from many tumours,

investigators can define a genomic region that is lost in all cases. The 'endgame' involves identifying the genes in this deleted DNA segment and screening human tumours for mutations in, or silencing of, the remaining copy of the candidate tumour-suppressor genes.

The discovery of these genes has been greatly facilitated by the availability of the human genome sequence, together with efficient DNA-sequencing technologies, and techniques such as high-density single-nucleotide-polymorphism arrays, which detect single-nucleotide variations within the population. Unfortunately, this general procedure becomes problematic when tumorigenesis results from inactivation of a single allele (haploinsufficiency). Indeed, it now seems that haploinsufficiency is a frequent genetic mechanism underlying human cancers<sup>4</sup>.

If discovering a 'classic' tumour-suppressor gene is like finding a needle in a haystack, the challenge involved in uncovering haploinsufficient tumour-suppressor genes is akin to finding a specific piece of hay in a haystack. This is because the traditional criterion for validating a tumour suppressor — mutations in both alleles — does not apply to haploinsufficient tumour-suppressor genes, as they retain one normal allele.

Several strategies have been used to address this formidable problem. One way is to look for cancer in animals that have inherited one mutant allele of a relevant gene. For example, studies of mice lacking the *p53* gene demonstrated<sup>5</sup> that inactivation of one or both alleles of this tumour-suppressor gene can promote tumorigenesis. Another way is to expose haploinsufficient mice and their normal littermates to chemical mutagens or radiation and to

compare the incidence and acceleration of tumour formation in the two sets of animals<sup>6</sup>. Yet another strategy is chromosome engineering, which involves producing a chromosome that lacks a large region of DNA<sup>7</sup>. An elegant example of this approach is a study<sup>8</sup> that identified *CHD5* as the elusive tumour-suppressor gene in chromosomal band 1p36.3, a region of DNA that is commonly deleted in human cancers.

Analysis of human cancers can also provide evidence for haploinsufficiency. For instance, monoallelic mutations in genes that encode various components of a B-lymphocyte differentiation pathway were identified through studies of acute lymphoblastic leukaemia in children<sup>9</sup>.

Ebert *et al.*<sup>2</sup> describe a creative new approach to the search for haploinsufficient tumour-suppressor genes that harnesses the technique of RNA interference. Using this technique, they systematically reduced the expression of each candidate tumour-suppressor gene associated with the 5q<sup>−</sup> syndrome. In this disorder, the commonly deleted segment of 5q31–5q33 spans about 1.5 megabases of DNA and includes 40 genes. Molecular investigation did not reveal mutations in the second allele of any candidate tumour-suppressor gene in this DNA segment, suggesting that the disease is caused by haploinsufficiency.

To determine which gene (or genes) might be involved in the 5q<sup>−</sup> syndrome, Ebert *et al.* synthesized several short, 'hairpin' RNA sequences that were complementary to each candidate gene. They then expressed these molecules in immature haematopoietic (CD34<sup>+</sup>) cells from normal bone marrow, and induced the cells to differentiate into precursors of red blood cells (erythroid cells) in culture.

The authors identify the haploinsufficient tumour-suppressor gene associated with the 5q<sup>−</sup> syndrome as *RPS14*. They validate this connection by showing that expressing *RPS14* in CD34<sup>+</sup> cells from patients with the 5q<sup>−</sup> syndrome enhances erythroid-cell differentiation and normalizes the activation level of genes specifically expressed in these red-blood-cell precursors. They also show that reducing *RPS14* expression in normal CD34<sup>+</sup> cells induces a gene-expression profile that correlates with responsiveness to the drug lenalidomide. Treatment with this drug results in loss of the abnormal population of 5q<sup>−</sup> cells and improvement of the anaemia in most 5q<sup>−</sup> syndrome patients. Together, the results provide strong evidence that *RPS14* functions as a haploinsufficient tumour-suppressor gene in the 5q<sup>−</sup> syndrome.

The protein encoded by *RPS14* is an essential component of the 40S subunit of a cellular organelle known as the ribosome, the site of protein synthesis. The *RPS14* protein is essential for efficient formation of the RNA–protein complexes involved. Ebert *et al.* find that ribosome synthesis in CD34<sup>+</sup> cells of 5q<sup>−</sup> syndrome patients is impaired. They also note that two

other ribosomal genes — *RPS19* and *RPS24* — are mutated in people with Diamond–Blackfan anaemia, a congenital form of anaemia that shares certain disease features with the 5q– syndrome<sup>10</sup>.

Several questions arise from these results. For example, how do reduced levels of *RPS14*, *RPS19* and *RPS24* proteins impair the formation of red blood cells? Are further mutations required for the 5q– syndrome to transform into acute myeloid leukaemia, and if so, what are they? Does *RPS14* haploinsufficiency contribute to the pathogenesis of other subtypes of myelodysplastic syndrome or acute myeloid leukaemia that are also associated with abnormalities in chromosome 5, perhaps by interacting with the effects of loss of genes on other regions of 5q? What are the molecular mechanisms underlying the dramatic genetic and clinical responses to lenalidomide in the 5q– syndrome, and why do some patients either fail to respond to this drug or relapse after an initial remission? And will treatment with lenalidomide or a related drug be beneficial in severe cases of Diamond–Blackfan anaemia?

It is also worth considering how the RNA-interference strategy developed by Ebert *et al.* might be extended to identify other haploinsufficient tumour-suppressor genes. In many respects, the 5q– syndrome is an optimal setting for using this approach — patients show consistent characteristics at a cellular level; the short hairpin RNA sequences used by the authors can readily be introduced into cultured immature bone-marrow cells; and there are established systems for monitoring cell survival and cell differentiation in liquid cultures. By contrast, deletions of chromosome bands 5q31, 7q22 and 20q12, which are found in many blood-related cancers, are frequently associated with other cytogenetic and molecular abnormalities that might influence the behaviour of cultured cells. Extending this approach to non-blood-related cancers poses yet other challenges, although these might be met by carefully investigating matched cell lines with or without deletions of a specific chromosomal segment. Despite the potential difficulties, however, the work of Ebert *et al.*<sup>2</sup> is a *tour de force* that holds great potential for addressing the problem of discovering and validating haploinsufficient tumour-suppressor genes. ■

Kevin M. Shannon is in the Department of Pediatrics and the Comprehensive Cancer Center, University of California, San Francisco, 513 Parnassus Avenue, San Francisco, California 94143-0519, USA.

Michelle M. Le Beau is in the Section of Hematology/Oncology and the Cancer Research Center, University of Chicago, 5841 South Maryland Avenue, MC2115, Chicago, Illinois 60637, USA.

e-mail: shannonk@peds.ucsf.edu

2. Ebert, B. L. *et al.* *Nature* **451**, 335–339 (2008).
3. Weinberg, R. A. *Science* **254**, 1138–1146 (1991).
4. Fodde, R. & Smits, R. *Science* **298**, 761–763 (2002).
5. Venkatchalam, S. *et al.* *EMBO J.* **17**, 4657–4667 (1998).
6. Joslin, J. M. *et al.* *Blood* **110**, 719–726 (2007).

7. Ramirez-Solis, R., Liu, P. & Bradley, A. *Nature* **378**, 720–724 (1995).
8. Bagchi, A. *et al.* *Cell* **128**, 459–475 (2007).
9. Mullighan, C. G. *et al.* *Nature* **446**, 758–764 (2007).
10. Gazda, H. T. & Sieff, C. A. *Br. J. Haematol.* **135**, 149–157 (2006).

## ASTRONOMY

# Elliptical view of galaxies past

Andrea Cimatti

**How and when galaxies assembled their mass to become the structures seen today are among astronomy's big outstanding questions.**

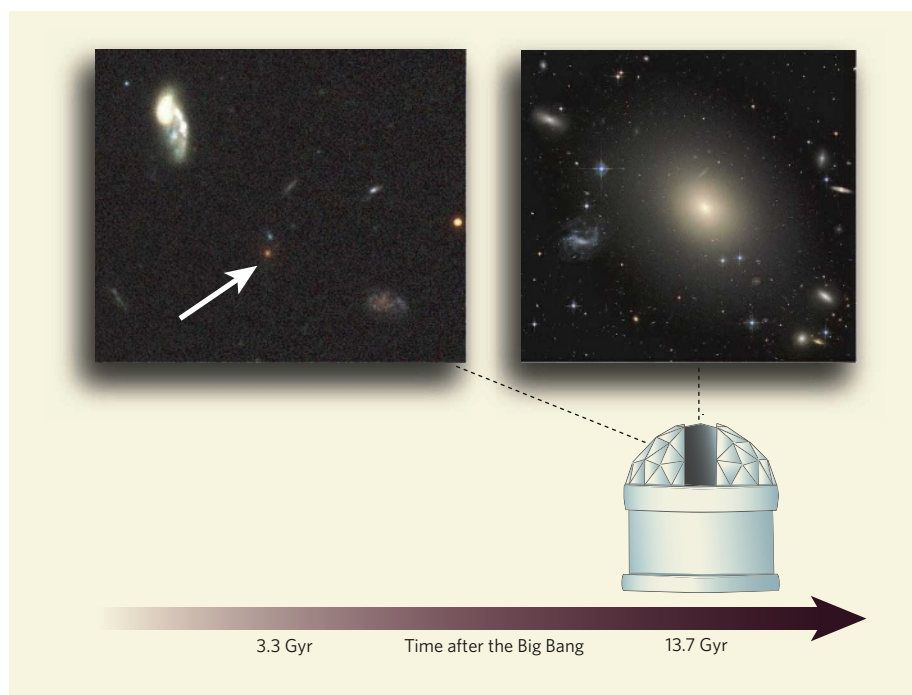
**A comprehensive study of nearby galaxies provides a new angle on the issue.**

Compared with other scientists, astronomers are at a disadvantage: they cannot perform laboratory experiments on stars and galaxies. But they can exploit a unique advantage: thanks to the finite speed of light, they can observe objects as they were in the past. Current telescopes allow galaxies to be observed as they were back to about 13 billion years ago. What a palaeontologist wouldn't give for a similar time machine for taking pictures of dinosaurs when they were alive!

Unfortunately, however, the direct study of astronomical objects at very great distances using the current generation of telescopes is fraught with difficulty, and the available

sample of such objects is still rather small. Writing in *The Astrophysical Journal*, Jimenez and colleagues<sup>1</sup> circumvent this problem by analysing the spectra of a very large sample of nearby (that is, present-day) 'early-type' galaxies to decode their history. Their results place tight constraints on the different evolutionary paths of galaxies as a function of their mass, providing a crucial reference for observational studies of distant galaxies and for theoretical models of galaxy formation.

In the currently favoured model of the Universe's evolution, galaxies formed gradually through hierarchical merging of 'haloes' of invisible dark matter<sup>2</sup>. The first galaxies are



**Figure 1 | Without looking back.** With the current generation of large telescopes, we can — just about — study the physical and evolutionary properties of distant elliptical galaxies when the Universe was just 3 billion years (Gyr) old. But the sample of galaxies at this distance is small. Jimenez *et al.*<sup>1</sup> adopt a different approach, in which they study in detail the properties of the elliptical galaxies in the present-day Universe and reconstruct their past evolution from the clues present in their spectra. Here, the tiny, compact red galaxy of the left-hand image (arrow) has become the large, diffuse elliptical galaxy of the right-hand image.

1. Giagounidis, A. A. *et al.* *Hematology* **9**, 271–277 (2004).

expected to have reached their final form and mass only rather recently. The early-type galaxies studied by Jimenez *et al.* are galaxies of 'elliptical' and 'lenticular' shape that have not formed features such as the characteristic arms of spiral galaxies. They contain most of the stellar mass in the present-day Universe, and so are the primary probes for investigating how galaxies assembled over cosmic time. But certain observations of early-type galaxies<sup>3</sup> seem to be in conflict with the hierarchical model. The number density of massive early-type galaxies, for example, is much the same now as it was 6 billion or 7 billion years ago<sup>4,5</sup>, whereas one would expect it to become less as galaxies merge. Old, massive early-type galaxies exist at even larger distances, representing a look-back time of some 10 billion years<sup>6</sup>. The Universe itself is about 13.7 billion years old: how was it possible to assemble these systems so rapidly when the Universe was so young?

Jimenez *et al.*<sup>1</sup> analysed the spectra of some 40,000 nearby early-type galaxies selected from the Sloan Digital Sky Survey<sup>7</sup> to find out. The spectrum of each galaxy carries a 'fossil record': a history of star formation and metal abundances in the galaxy. It can be reconstructed by analysing the shape of the spectrum of thermal radiation emitted by the stars in the galaxy and the absorption lines in it. The authors find that the evolution of early-type galaxies is characterized by a strong 'downsizing' effect<sup>8,9</sup>: massive galaxies form most of their stars and

stellar mass faster and earlier (more than 10 billion years ago) than do low-mass galaxies. Star formation in massive early-type galaxies was rapidly suppressed early in the galaxies' formation by the action of supernovae (exploding stars) and/or an active galactic nucleus (a supermassive black hole at the galaxy's centre). These objects heated the gas and prevented its collapsing further to form new stars. The abundance of metal elements evolves in proportion to the galaxy's mass: gas is trapped for longer in the deeper potential wells of more massive galaxies, and is consequently more enriched in the metals produced during the processes of star formation.

These results strengthen previous studies<sup>10</sup> and are crucial for two main reasons. First, they provide a statistically solid reference work that both theoretical and observational studies of galaxy formation should take into account. Second, they make clear predictions with which observations of early-type galaxies at large look-back times can be compared, as more of these become available. Reassuringly, early-type galaxies identified so far from when the Universe was between 2 billion and 3 billion years old<sup>6</sup> do indeed have the properties that Jimenez and colleagues' study<sup>1</sup> of their present-day counterparts leads us to expect. The clear implication is that most of the star formation and mass assembly of massive early-type galaxies took place during the first 2 billion to 3 billion years after the Big Bang.

The two complementary approaches of fossil-record analysis and look-back time studies are finally providing a coherent answer to the long-standing question of massive-galaxy formation. The next generation of telescopes, such as the European Space Agency's Herschel, the European-American Atacama Large Millimeter Array and NASA's James Webb Space Telescope, will allow the direct study of larger samples of distant galaxies. Once we can unambiguously identify the star-forming precursors of present-day early-type galaxies at earlier cosmic times, we shall be able to start understanding the physical and dynamical processes that drove their formation and shaped their structure. ■

Andrea Cimatti is in the Dipartimento di Astronomia, Alma Mater Studiorum, Università di Bologna, Via Ranzani 1, I-40127 Bologna, Italy. e-mail: a.cimatti@unibo.it

1. Jimenez, R., Bernardi, M., Haiman, Z., Panter, B. & Heavens, A. F. *Astrophys. J.* **669**, 947–951 (2007).
2. Springel, V. *et al. Nature* **435**, 629–636 (2005).
3. Renzini, A. *Annu. Rev. Astron. Astrophys.* **44**, 141–192 (2006).
4. Cimatti, A., Daddi, E. & Renzini, A. *Astron. Astrophys.* **453**, L29–L33 (2006).
5. Bundy, K., Treu, T. & Ellis, R. S. E. *Astrophys. J.* **665**, L5–L8 (2007).
6. Cimatti, A. *et al. Nature* **430**, 184–187 (2004).
7. [www.sdss.org](http://www.sdss.org)
8. Cowie, L., Songaila, A., Hu, E. M. & Cohen, J. G. *Astron. J.* **112**, 839–864 (1996).
9. Gavazzi, G. & Scoddeggio, M. *Astron. Astrophys.* **312**, L29–L32 (1996).
10. Thomas, D., Maraston, C., Bender, R. & Mendes de Oliveira, C. *Astrophys. J.* **621**, 673–694 (2005).

## IMMUNOLOGY

# Cascade into clarity

Fayyaz S. Sutterwala and Richard A. Flavell

**Immune mediator molecules such as antimicrobial peptides are crucial for host responses to pathogens. Akirins are the latest identified components of a signalling cascade that leads to these responses in insects and mice.**

The availability of powerful genetic tools to study the fruitfly *Drosophila melanogaster*, and the striking similarities of this insect's immune system to that of mammals, makes *Drosophila* a valuable organism for researchers interested in innate (nonspecific) immune responses. Indeed, among other advances, the discovery of Toll-like receptors, which are essential mediators of innate immunity in mammals, came about through studies in *Drosophila*. Reporting in *Nature Immunology*, Goto *et al.*<sup>1</sup> have used this insect to identify another essential player in the innate immune system that is structurally highly conserved in mammals. This gene, which the authors named *Akirin* — after the Japanese phrase '*akiraka ni suru*', which means 'making things clear' — encodes a nuclear protein that affects the transcription of genes regulated by the transcription factor known as NF- $\kappa$ B, which

is found in almost all animal cells.

When an organism suffers a microbial infection, its immune system rapidly mounts a defence characterized by the production of large amounts of cytokines and antimicrobial peptides. This innate response is mediated by pattern-recognition receptors, including Toll-like receptors, that detect evolutionarily conserved structures, such as peptidoglycan subunits, associated with pathogens.

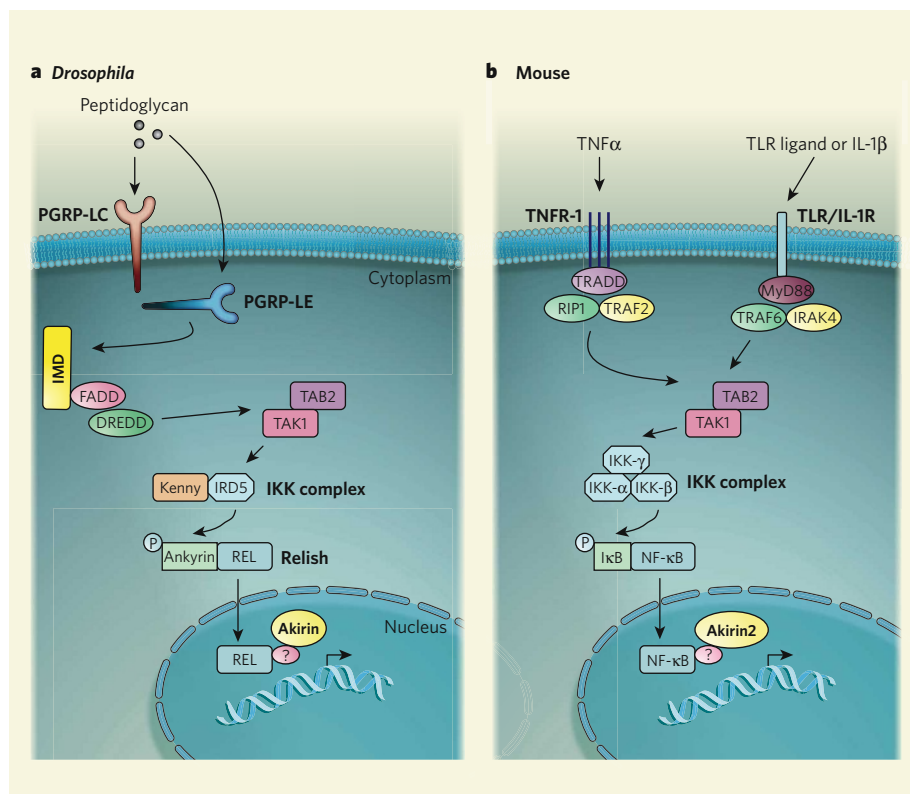
In *Drosophila*, two main pathways lead to the production of antimicrobial peptides: the Toll pathway and the immune deficiency (Imd) pathway. The Toll pathway responds to Gram-positive bacteria and fungal pathogens<sup>2,3</sup>, whereas the Imd pathway, in which Akirin plays a crucial part, is turned on in response to infections with Gram-negative bacteria<sup>4–6</sup>.

The Imd signalling cascade culminates in the activation of Relish, an NF- $\kappa$ B-like

transcription factor<sup>7</sup> (Fig. 1a). Initially, the binding of peptidoglycan subunits of Gram-negative bacteria to the *Drosophila* peptidoglycan-recognition proteins PGRP-LC and PGRP-LE activates the Imd protein. Active Imd recruits the 'death-related proteins' Fadd and Dredd, which in turn activate a complex of TAK1 and TAB2 proteins. Further down the pathway, two enzymes, IRD5 (the homologue of mammalian IKK- $\beta$ , which is involved in NF- $\kappa$ B activation) and Kenny (the homologue of mammalian IKK- $\gamma$ ), are activated. These enzymes add a phosphate group to Relish, thus marking it for cleavage. Relish then moves to the nucleus, where it drives the transcription of genes encoding antimicrobial peptides.

Goto and colleagues<sup>1</sup> now show that this story is incomplete. Using the technique of RNA interference in *Drosophila* S2 cells, they show that, in response to infection with Gram-negative bacteria, Akirin is required for the expression of the antimicrobial peptide Attacin, which is an essential end-product of the Imd pathway. This observation is unexpected because, apart from a nuclear-localization signal, Akirin has none of the identifiable structural domains characteristic of signalling molecules.

The authors then used genetic-interaction studies to show that Akirin functions downstream of, or at the same level as, Relish (Fig. 1a).



**Figure 1 | Role of Akirin proteins in producing immune mediators following microbial infection.** **a**, In *Drosophila*, peptidoglycan components of Gram-negative bacteria activate the Imd pathway, which results in the movement of Relish (an NF- $\kappa$ B-like transcription factor) to the nucleus. Relish then mediates transcription of genes that encode antimicrobial peptides. Goto *et al.*<sup>1</sup> identify Akirin, a nuclear factor that acts late in this signalling cascade and is required for Relish-mediated gene transcription. **b**, In mammals, the activation of the TNF receptor 1 (TNFR-1), Toll-like receptor (TLR) or interleukin-1 receptor (IL-1R) turns on a signalling cascade that results in movement of NF- $\kappa$ B to the nucleus and activation of gene transcription. The authors find that, in mice, a structurally highly conserved homologue of *Drosophila* Akirin, Akirin2, is required for NF- $\kappa$ B-mediated gene transcription.

Moreover, they found that Akirin deficiency does not affect the Toll pathway, suggesting that this protein is involved in the production of antimicrobial peptides only through the Imd pathway. Consistent with these *in vitro* findings, reducing Akirin levels in live flies using RNA interference increased the flies' susceptibility to infection with Gram-negative bacteria. These findings clearly establish Akirin's role in the Imd signalling pathway. But this protein probably has other functions too. Goto *et al.* show that mutant flies lacking the *Akirin* gene are not viable, implying a crucial role in *Drosophila* embryonic development.

Does Akirin have a similar function in mammals? In looking at this question, the authors find that structurally highly conserved *Akirin* is present in mice as two homologues (*Akirin1* and *Akirin2*). To investigate the function of mammalian Akirins, they generated mice deficient in either *Akirin1* or *Akirin2*. Neither *Akirin1*-deficient mice nor cells derived from these animals have any obvious unusual characteristics. However, the function of *Akirin1* could be hidden through functional redundancy in the presence of *Akirin2*, a point that requires further investigation.

Like *Akirin* in *Drosophila*, *Akirin2* is required for embryonic development, and Goto *et al.* found that mice lacking this gene die by embryonic day 9.5. Fibroblast cells derived from *Akirin2*-deficient mouse embryos showed selective defects in NF- $\kappa$ B-dependent gene expression following stimulation through pathways involving the Toll-like receptor, interleukin-1 receptor or TNF receptor. All of these pathways converge on the activation of the mammalian TAB2-TAK1 complex, which in turn activates

the IKK complex. Through phosphorylation, the active IKK complex causes the degradation of the NF- $\kappa$ B inhibitor I $\kappa$ B, allowing NF- $\kappa$ B to enter the nucleus (Fig. 1b). The authors postulate that, like *Drosophila* Akirin, which acts downstream of Relish, Akirin2 functions downstream of NF- $\kappa$ B.

How do Akirins regulate gene transcription in the nucleus? Although preliminary studies failed to show a direct interaction of Akirins with DNA or with Relish, it is possible that they interact with an intermediary molecule that then engages with DNA and/or Relish, or is otherwise involved in transcription. It is also likely that Akirins are involved in regulating transcription factors other than NF- $\kappa$ B. The fact that Akirin is a potential modulator of the Wnt-Wingless developmental pathway in *Drosophila*<sup>8</sup> suggests that it might regulate the associated  $\beta$ -catenin transcription factor. Similarly, Akirin could be involved in regulating the GATA transcription factor, as it interacts with the GATA-related protein pannier, which is essential for thorax development in *Drosophila*<sup>9</sup>.

A clear picture emerges: the functions of Akirins probably extend beyond the immune system, as do those of many other genes involved in immunity, and which also have roles in development. The *toll* gene, for example, which is essential for innate immune responses in *Drosophila*, was first identified as a developmental gene. So the results of Goto *et al.* have opened avenues of research that not only may help to unravel the complexities of the inflammatory signalling pathway in which Akirins function, but also may aid our understanding of the function of these

molecules in embryonic development.

Fayyaz S. Sutterwala is in the Department of Medicine, Inflammation Program, University of Iowa, Iowa City, Iowa 52241, USA. Richard A. Flavell is in the Department of Immunobiology and the Howard Hughes Medical Institute, Yale University School of Medicine, New Haven, Connecticut 06520, USA.  
e-mails: fayyaz-sutterwala@uiowa.edu;  
richard.flavell@yale.edu

- Goto, A. *et al.* *Nature Immunol.* **9**, 97–104 (2008).
- Lemaitre, B., Nicolas, E., Michaut, L., Reichhart, J. M. & Hoffmann, J. A. *Cell* **86**, 973–983 (1996).
- Michel, T., Reichhart, J., Hoffmann, J. A. & Royet, J. *Nature* **414**, 756–759 (2001).
- Gottar, M. *et al.* *Nature* **416**, 640–644 (2002).
- Ramet, M., Manfrulli, P., Pearson, A., Mathey-Prevot, B. & Ezekowitz, R. A. *Nature* **416**, 644–648 (2002).
- Choe, K. M., Werner, T., Stoven, S., Hultmark, D. & Anderson, K. V. *Science* **296**, 359–362 (2002).
- Ferrandon, D., Imler, J. L., Hetru, C. & Hoffmann, J. A. *Nature Rev. Immunol.* **7**, 862–874 (2007).
- DasGupta, R., Kaykas, A., Moon, R. T. & Perrimon, N. *Science* **308**, 826–833 (2005).
- Peña-Rangel, M. T., Rodriguez, I. & Riesgo-Escovar, J. R. *Genetics* **160**, 1035–1050 (2002).

#### Correction

In the News & Views article on thermoelectric silicon nanowires "Materials science: Desperately seeking silicon" by Cronin B. Vining (*Nature* **451**, 132–133; 2008), we unfortunately swapped the contexts in which the experiments in the two papers concerned were conducted. Hochbaum *et al.* (reference 4 of the article) suspended their nanowires above a silicon substrate, whereas those of Boukai *et al.* (reference 3) were supported on a thin silica platform that was fully suspended in a vacuum. The nanowire cross-sections of Hochbaum *et al.* were also not perfectly circular, but irregularly shaped, with diameters between 20 and 300 nm.

## SOLID-STATE PHYSICS

## Join the dots

Galina Khitrova and H. M. Gibbs

**A new variation on an old theme in atomic physics, a spectral distortion known as the Fano effect, has been revealed — not in an atom, but in an artificial nanostructure known as a quantum dot.**

The Fano effect is a quantum-mechanical interference phenomenon characterized by an asymmetrical broadening of spectral lines that pops up all over the place when certain materials absorb light. In 1981, it was predicted<sup>1</sup> that using light from a strong resonant laser beam would completely alter the spectrum of the Fano effect. That prediction has still not been fulfilled; but on page 311 of this issue, Kroner *et al.*<sup>2</sup> describe how using a resonant laser beam reveals a Fano effect that had hitherto remained obscured.

Rather than using atoms, as has generally been the case in investigations of the Fano effect, the authors' demonstration uses single quantum dots. These nanoscale semiconducting structures are being used not only to study the fundamental interactions between photons and systems of energy levels similar to those in atoms, but also for constructing minuscule light-emitting diodes and lasers. Kroner and colleagues' Fano effect could have practical implications because it represents a sensitive way to detect the coupling of a transition between two energy levels to a continuum of energy states. Such couplings are usually undesirable for applications using quantum dots.

Kroner and colleagues' quantum dots are made of the semiconductor indium arsenide (InAs) capped with a thin layer of gallium arsenide (GaAs). When light is shone on one of these quantum dots, the absorption of a photon excites an electron out of the semiconductor's valence band and into its conduction band. This excitation produces not only an electron, but also a hole, equivalent to a positive charge, where the electron used to be. Both the electron and the hole are tightly confined in three dimensions within the dot. Their energies are thus quantized into a set of discrete levels, just as in an atom.

By adjusting the thickness of the GaAs capping layer and applying an electric field, the authors also created an effective quantum well, which contains a two-dimensional continuum of energy states for holes at energies overlapping the discrete energy of the hole, but spatially separated by a thin barrier. A hole, generated along with an electron by absorption of an incident photon, can tunnel into this well. By measuring the absorption spectrum of an individual quantum dot, Kroner *et al.* looked for Fano interference between these two ways of absorbing a photon: the usual transition producing an electron and a hole in discrete energy levels, and the much weaker

discrete–continuum transition through the quantum tunnel.

With a very weak laser beam, the authors saw no hint of the weak tunnel coupling. But as they increased the laser power, the discrete–discrete absorption decreased towards the level of the discrete–continuum transition<sup>2</sup>. Interference between these two pathways when they are of almost equal strength causes the absorption spectrum to take on the asymmetrical shape characteristic of a strong Fano effect. The credibility of this interpretation is strengthened by the fact that the asymmetry in the spectrum disappears if the continuum state is removed by making the capping layer thinner.

So what? The important point to bear in mind is that the complete isolation of discrete–discrete transitions in quantum dots is essential for almost all fundamental experiments on single quantum dots. This isolation could be spoiled by a small coupling to an unknown continuum. By driving the discrete–discrete transition with a continuous-wave laser, a relatively weak leak to an unwanted continuum can be detected through the clear signal of Fano distortion. The effect will thus be a useful diagnostic tool in designing quantum-dot structures to eliminate such effects.

This research is the latest in a long progression ever since it was first proposed in 1970<sup>3</sup> that man-made quantum structures could be designed that would mimic the quantized energy levels of an atom's potential well. The experimental breakthrough came with the development of the technique known as molecular-beam epitaxy, which allows single layers of semiconductor materials to be grown one on top of each other. Quantized energy levels were soon observed<sup>4</sup> in quantum energy wells produced by growing GaAs between potential barriers consisting of the closely related semiconductor aluminium gallium arsenide (AlGaAs).

A few years later, Alexei Ekimov hypothesized that the losses in optical fibres that were then preventing their use for telecommunications were the result of semiconductor impurities. He introduced controlled amounts of semiconductor compounds into glass to test that theory. Ekimov noticed bumps in the absorption spectra of the glass that became more widely separated in energy as the volumes of the semiconductor regions were reduced. By analogy with the quantum-well phenomenon, he concluded<sup>5</sup> that this was a signature of three-dimensional quantum confinement<sup>6</sup>. The quantum dot had arrived.

Quantum dots in glass and in colloidal solutions are useful for some applications. But it was obvious that the ability to grow dots within the easily doped heterostructures that dominate the world of semiconductor light-emitters would be highly desirable. This would enable the production of quantum-dot lasers that would require a reduced threshold current and maintain greater wavelength stability against temperature changes. The development<sup>7</sup> of self-organization techniques that use mechanical strain to trick the usual planar growth of molecular-beam epitaxy into becoming three-dimensional has permitted the control of quantum-dot density, diameter and height. High dot densities are ideal for lasers of very small volume. Low dot densities allow the isolation of a single quantum dot, providing sources of single photons on demand and quantum entangled states for quantum information science<sup>8</sup>.

In all this, it is curious and instructive to note the mutual benefits of basic and applied research. An applied goal, reducing losses in optical fibres, led to the fundamental discovery of quantum dots; the applied goal of growing quantum dots for lasers resulted in dots that now compete with atoms for use in basic research. A quantum dot has the distinct advantage over an atom of being nailed to one place, and not needing multiple highly stabilized laser beams to trap it; the dot structure is also monolithic, tiny and long-lived.

Indeed, quantum dots have arguably already become more useful than atoms in a number of instances, such as an efficient source of single photons on demand<sup>9</sup>. Kroner and colleagues' use<sup>2</sup> of the nonlinear Fano effect as, in essence, a tremendous sensitivity amplifier for the spectroscopic identification of weak continuum spectra adds another instance to the growing list. ■

Galina Khitrova and H. M. Gibbs are in the College of Optical Sciences, University of Arizona, Tucson, Arizona 85721, USA.  
e-mail: galina@optics.arizona.edu

1. Rzazewski, K. & Eberly, J. H. *Phys. Rev. Lett.* **47**, 408–412 (1981).
2. Kroner, M. *et al.* *Nature* **451**, 311–314 (2008).
3. Esaki, L. & Tsu, R. *IBM J. Res. Dev.* **14**, 61–65 (1970).
4. Dingle, R., Wiegmann, W. & Henry, C. H. *Phys. Rev. Lett.* **33**, 827–830 (1974).
5. Ekimov, A. I. & Onushchenko, A. A. *JETP Lett.* **34**, 345–349 (1981).
6. Ekimov, A. I., Efros, A. L. & Onushchenko, A. A. *Solid State Commun.* **56**, 921–924 (1985).
7. Petroff, P. M., Lorke, A. & Imamoglu, A. *Phys. Today* **54** (5), 46–52 (2001).
8. Khitrova, G., Gibbs, H. M., Kira, M., Koch, S. W. & Scherer, A. *Nature Phys.* **2**, 81–90 (2006).
9. Pelton, M. *et al.* *Phys. Rev. Lett.* **89**, 233602 (2002).

**Cover illustration**

Pinnacles eroded from sedimentary rock, with melting snow, in Bryce Canyon National Park, Utah. (Courtesy of T. Dempsey/Photoseek.com)

**Editor, *Nature***

Philip Campbell

**Insights Publisher**

Sarah Greaves

**Publishing Assistant**

Claudia Banks

**Insights Editor**

Karl Ziemelis

**Production Editor**

Davina Dadley-Moore

**Senior Art Editor**

Martin Harrison

**Art Editor**

Nik Spencer

**Sponsorship**

Emma Green

**Production**

Jocelyn Hilton

**Marketing**

Katy Dunningham

Elena Woodstock

**Editorial Assistant**

Alison McGill

# YEAR OF PLANET EARTH

**A**s we progress into the twenty-first century, modern society faces one of its greatest challenges — climate change. Earth scientists are uniquely placed to help tackle this issue, as well as to help society reduce the risks from natural hazards and use Earth's resources sustainably.

To achieve these goals, it is essential that Earth scientists and society interact in mutually beneficial ways, as Ted Nield and Frank Press reflect in the essays that open and close this collection. But it is also crucial that Earth scientists are excited and inspired by science in its own right, and it is this aim that we hope to fulfil through the other articles in this supplement. These informal, sometimes opinionated, pieces look back at recent developments in the Earth sciences and consider where future advances might lie.

These ideas have much in common with the philosophy behind the International Year of Planet Earth, a joint initiative by the United Nations Educational, Scientific and Cultural Organization (UNESCO) and the International Union of Geological Sciences. This project aims to capture people's imagination with the knowledge accumulated by Earth scientists and to ensure that this information is used to benefit society, and we hope that this supplement will contribute to these goals.

With *Nature Geoscience*, Nature Publishing Group has just launched a new journal that also supports the goals of the International Year of Planet Earth. Alongside *Nature*, *Nature Geoscience* will publish research, commentary and analysis across the entire spectrum of the Earth sciences.

We are pleased to acknowledge the financial support of the International Year of Planet Earth (IYPE) and the International Union of Geological Sciences in producing this supplement. As always, *Nature* carries sole responsibility for all editorial content.

Joanna Thorpe, Associate Editor,  
Juliane Mössinger and John VanDecar, Senior Editors

**ESSAY****258 A tribe of jobbing ditchers**

T. Nield

**FEATURES****261 A planetary perspective on the deep Earth**

D. J. Stevenson

**266 Using seismic waves to image Earth's internal structure**

B. Romanowicz

**269 Mineralogy at the extremes**

T. S. Duffy

**271 Earthquake physics and real-time seismology**

H. Kanamori

**274 From landscapes into geological history**

P. A. Allen

**277 The rise of atmospheric oxygen**

L. R. Kump

**279 An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics**

J. C. Zachos, G. R. Dickens &

R. E. Zeebe

**284 Unlocking the mysteries of the ice ages**

M. E. Raymo & P. Huybers

**286 Ocean circulation in a warming climate**

J. R. Toggweiler & J. Russell

**289 Terrestrial ecosystem carbon dynamics and climate feedbacks**

M. Heimann & M. Reichstein

**293 An Earth-system perspective of the global nitrogen cycle**

N. Gruber & J. N. Galloway

**297 A steep road to climate stabilization**

P. Friedlingstein

**299 Small-scale cloud processes and climate**

M. B. Baker & T. Peter

**ESSAY****301 Earth science and society**

F. Press

# A tribe of jobbing ditchers

Ted Nield

**Earth science, a field in which science and profession have been intimately linked, has grown through the practicalities imposed by industrialization and war but must now revamp to address climate change.**

The celebrated English engineer and entrepreneur Matthew Boulton had a low opinion of the emerging profession of canal engineer. But for all this, the despised tribe would soon include the great William Smith, 'the father of English geology'. By 1799, Smith had spent almost six years as a 'jobbing ditcher', laying out the Somerset Coal Canal and superintending its building. In that year, he took a circular map of the district of Bath and shaded the different rock types (now widely cited as the world's first geological map), and similarly coloured the geology of England on a small-scale map. This map of England was the precursor of his great geological map of 1815, known (since Simon Winchester's best-selling book) as "the map that changed the world".

Smith's ditching activities were the ideal experiment — digging a continuous trench through gently dipping fossil-rich rocks — enabling him to prove a hypothesis that he had been forming since earlier work in the coal mines of Somerset. This theory held that all sediments of the same age would carry the same fossils. This became the 'law of the identification of strata by contained fossils', which — combined with the 'law of superposition' (old stuff is on the bottom, and young stuff is on the top) — enabled Smith to identify rocks of the same relative age. He then coloured their outcrop patterns on a topographic map from one side of the country to the other. Stratigraphic mapping was born.

But why was it left to the son of a Somerset blacksmith with little or no formal education past the age of 11 to come up with these simple but powerful ideas? The world had hardly been short of geological savants in preceding centuries. In 1807, some years before Smith finally published his great map and fossil album, the Geological Society of London had been founded. With an admirable disregard for superstition, 13 men put their names to its founding document after a meeting on Friday, 13 November. They inaugurated a society dedicated to observation and objective description and eschewed airy and overambitious 'theories of the Earth' that they felt (along with their French contemporaries Georges Cuvier and Alexandre Brongniart) had bedevilled attempts to understand the Earth's deep history.

Yet the society's much-imitated commitment to objective description within a stated research agenda, although intellectually down to earth, was still far from being 'applied science'. Many of the founders — and those who later joined the growing society — may have been practical men, but they lacked one vital spur. They either had money or earned their living elsewhere. They were gentlemen, after all.

To give them credit, when Smith exhibited his map, they recognized its worth, bought a copy and, in modern parlance, plagiarized it (using it uncredited as the basis for their own, much improved, map of England and Wales). But this act, which seems unbelievably callous today, did not mean that they were immoral. Their interaction with Smith was dictated by the class gulf between them, and the concept of intellectual property was in its infancy. They had paid him for his labour — what more did a man of his class expect? To be sure, it was not long before they began to feel embarrassed about this, and eight years before his death, they presented Smith with the first Wollaston Medal, the society's highest honour.

## Theory and practice

Smith received his apology in the form of the Wollaston Medal, but his real revenge was that he, and not the society gentlemen, was granted the honour of the breakthrough. Smith was pioneering a profession, as well as a science. And, to any applied geologist, the idea of a geological map is almost self-evident. That is, before you can do anything, you need to know what rocks lie where, and what they are like.

This brings us to a central fact about scientific geology: its essential practicality. Everything we need — all raw materials and nearly all energy — comes from the planet. This means that a geoscientist needs first to find these things and then to extract them economically. Human societies today simply could not survive without geology. It is therefore no surprise that the intellectual revolution that was the emergence of scientific geology was tied to industrial revolution. If industrialization had been more advanced in France than the United Kingdom, then the history of geoscience would undoubtedly be much less anglocentric than it is.

Geology is thought of, not quite wrongly, as belonging to the Victorian era. Victorian battles over evolution and the age of Earth are the stuff of modern legend. Charles Darwin and Thomas Huxley were both geologists — Darwin, by his own admission, was a geologist primarily, and Huxley, secondarily (although only Huxley, co-founder of the journal *Nature*, became president of the Geological Society of London). It is clear that the public correctly senses geology's inherent congruence with industry, manufacturing and empire.

The British Empire sent trade, military and scientific envoys across the globe. Darwin's eyes were opened as the gentleman-naturalist companion to the captain of HMS *Beagle*. The man who would become his champion, Huxley, sailed aboard a leaky frigate called HMS *Rattlesnake*. Scientists felt the spur of the British Empire as they accompanied such Royal Navy expeditions bent on creating accurate charts for trade and defence.

The biogeographer Philip Lutley Sclater was another such traveller, perhaps less known today. In his journeys, he saw that lemurs had a scattered geographical distribution that did not make sense. Lemurs might have crossed from Africa to Madagascar on rafts of vegetation, but was it probable that they had crossed all the way to Sri Lanka or the Malay Archipelago? No, there must once have been land between. Sclater had (although he never knew it) found early evidence for continental drift and was glimpsing Gondwanaland, the ancient southern lobe of Pangaea.

Or take the brothers William and Henry Blanford (born in a London house that would become Charles Dickens's editorial office), who in 1856 were recruited by Thomas Oldham to the nascent Geological Survey of India. They quickly discovered mysterious glacial deposits near Cuttack, only a few degrees from the modern Equator. Similar rocks were soon found on all the parts of Gondwanaland, including Australia, South Africa and Antarctica. How glaciers could have extended over an entire hemisphere, mostly occupied by ocean, puzzled geologists for decades. Like Sclater's lemurs, these rocks were too similar to be so far apart.



"The map that changed the world":  
William Smith's great map of 1815.

By contrast, at the same time Alfred Russel Wallace, who was travelling to feed the appetite for exotic beasts, discovered animal species that were too different to be so close together. Why was the fauna of Eurasia suddenly replaced by that of Australasia across the narrow strait between Bali and Lombok? An invisible line between these islands delineated two great faunal realms. How was this so? The correct answer was the same — lateral motion of continents. But all these facts, won by empire and trade, would lie waiting through two world wars before scientists would get the tools needed to understand them.

### Rules of law

Rules, names and boundaries are effective means of colonization, and the Victorians mapped and codified everything they could get their hands on. Charting the world and naming its sounds and mountains are acts of possession, which efface indigenous history. Victorian geologists, for their part, set about conquering and colonizing the past. The Cambrian, Ordovician, Silurian and Devonian periods were all named after localities in the United Kingdom (or their pre-Roman inhabitants). Although the Permian was named after the Russian city of Perm, it had been identified by Roderick Impey Murchison, on his imperially sponsored 'geologizing' campaign across Russia. The names of the Tertiary epochs — Palaeocene, Eocene, Miocene and so on — were coined by University of Cambridge polymath William Whewell. 'Carboniferous' was just a fancy way of saying 'coal measures'. Barring a few French interlopers (Jurassic and Cretaceous), the British parcelling of time was effectively a result of imperialism.

### War footings

Imperial concerns provided the world with the main driver for geological exploration throughout much of the twentieth century — the oil and gas industry. Until the British Empire's trade was threatened by the rise of imperial Germany, oil had been a cottage industry. It was Winston Churchill who put it on a war footing and set it on the road to greatness, during the First World War.

The great dreadnought battleships were coal powered, because other fuels would have needed to be sourced abroad. But the disadvantages were starting to outweigh the advantages. Coaling could only be done in port; it was filthy, exhausting and required huge numbers of stokers. Oil, by contrast, had a larger calorific value. Ships could travel farther and faster on smaller boilers, and they could refuel at sea. The Royal Navy needed more efficient ships, and that was that.

Where was the oil to come from? The British government sent a delegation to the Gulf. Two companies took the lead: the Anglo-Dutch company Royal Dutch/Shell, and the Anglo-Persian Oil Company, a much smaller firm that was the forerunner of BP. Eventually, the UK government took a 51% share in the Anglo-Persian Oil Company and appointed two members to the board — so began the industrial-military complex. Geology is, of course, vital to war — sediment sampling in preparation for the Normandy landings is a famous example of heroism in the Second World War. But the two world wars, through the boost they gave to the British oil industry, did more for Earth science than anything else had done.

The benefits of bringing the resources of a cash-rich, highly capitalized industry to bear on geoscientific problems cannot be overestimated. The cutting edge of geoscientific thinking moved towards industry, which — with its facilities and intellectually adventurous environment — drew many of the best brains out of academia. Oil companies

could not help but become geological institutes. As Wallace Pratt, a founder of the American Association of Petroleum Geologists (AAPG), was soon to say, "Oil is found in the minds of men".

The AAPG — currently the world's largest professional geological society — also backed Alfred Wegener's hypothesis of continental drift, at a time when the United States was the greatest bastion of anti-continental-drift thought. In 1926, Willem van Waterschoot van der Gracht, who had left Europe after being dismissed by Shell and was also an AAPG founder, convened a scientific meeting to discuss continental drift, putting the fledgling organization's reputation on the line. On North American shores, however, van der Gracht found himself a lone 'drifter'. This European idea was almost universally condemned, to the extent that van der Gracht had to commission extra 'pro-drift' contributions from people who were not at the meeting and had to write a pro-drift commentary that took up 43% of the published volume. This report marked the establishment of a beachhead of progressive thought in the United States, with immense implications for hydrocarbon exploration, and it eventually paved the way for the reality of continental drift to be confirmed by the geophysicists who had formerly been most vociferously against it.

The conversion of geophysicists to continental drift came about because, as a result of the Second World War, they discovered the most convincing proof that any scientist can find — evidence from their own field. Suddenly, geophysical objections (which had largely centred on the assertion that there was no adequate mechanism) evaporated. During the First World War, the picture of the topography of the sea floor had been greatly improved by the introduction of echo-sounding devices. The ruggedness of the sea floor came as a surprise, as did the continuity of the Mid-Atlantic Ridge. But greater surprises lay in store. Magnetic surveys carried out in the years after the Second World War, using magnetometers adapted from airborne submarine detectors, began to find magnetic variations. It was these 'zebra stripes', symmetrically positioned across the axis of the worldwide mid-ocean ridge system, that finally convinced almost all geoscientists that the oceans were young and expanding (F. J. Vine and D. H. Matthews *Nature* **199**, 947–949; 1963). Continental drift became plate tectonics. Barely 150 years after the formation of the Geological Society of London, the much-despised ditchers had arrived at the Grand Unifying Theory of their field.

### Science and profession

The coincidence of a rash of unifying events in 2007–2008 — the United Nations International Year of Planet Earth, International Heliophysical Year, International Polar Year, Electronic Geophysical Year and the Geological Society of London's 200th birthday — provides opportunities for Earth scientists, both academic and professional, to see clearly where they must go and to speak with one voice. The reason for urgency is stark. Geoscientists have a unique understanding of Earth as a unified system of interacting components — the Earth system — which they must communicate. In the new battle against global climate change, geoscientists will fail in their duty to their fellow citizens if they fail in this. Practice and theory owe each other an equal debt. Each has provided grit to the other's oyster for 200 years. They must continue to do so, as geoscientists move on from the imperial reductionist past, apply the new holistic understanding of the Earth system, and have a proper role in the stewardship of a planet that humans cannot live without.

Ted Nield is editor of *Geoscientist*, the magazine of the Geological Society of London, and chair of the Outreach Programme Committee for the International Year of Planet Earth.

**Author Information** Reprints and permissions information is available at [npg.nature.com/reprints](http://npg.nature.com/reprints). Correspondence should be addressed to the author ([ted.nield@geolsoc.org.uk](mailto:ted.nield@geolsoc.org.uk)).

**The father of English geology, William Smith.**

# A planetary perspective on the deep Earth

David J. Stevenson

**Earth's composition, evolution and structure are in part a legacy of provenance (where it happened to form) and chance (the stochastics of that formation).**

Earth is an engine, tending to obliterate some of the evidence of events that are distant in time, but a memory is retained in its chemistry, its isotopes, the presence of the Moon, perhaps also in geophysical observables such as the temperature of the core and the nature of the mantle immediately above the core, and maybe even in the existence of plate tectonics and life. The remarkable growth in the study and understanding of Earth has happened in parallel with a spectacular era of planetary exploration, relevant astronomical discoveries and computational and theoretical advances, all of which help us to place Earth and its interior in a perspective that integrates the Earth sciences with extraterrestrial studies and basic sciences such as condensed-matter physics. However, progress on the biggest challenges in understanding the deep Earth continues to rely mainly on looking down rather than looking up.

## A planetary perspective

Earth is a planet — one of many. There is nothing particularly remarkable about our home, except perhaps that it is suitable for life like us — arguably a tautology. It happens to be the largest of its type in the Solar System, but as there are only three others of the terrestrial type (Mercury, Venus and Mars) this is not particularly significant. Among planets in general, it is small.

In the past decade, we have seen an astonishing explosion in our catalogue of planets outside the Solar System to about 250 so far (see the Extrasolar Planets Encyclopaedia, <http://exoplanet.eu/catalog.php>). These are mostly planets that we suspect are like Jupiter, very different from Earth. But as time goes on and detection methods improve, we can expect to find bodies that are Earth-like at least to the extent of being made predominantly of rock and iron, the primary constituents of our planet. Some would claim we might already be finding such bodies<sup>1</sup>, initially those that are more massive than Earth.

If planets were like atoms or molecules, or even crystals, we could speak of their characteristics (their DNA, so to speak) in a very compact way, just as a handbook might list the properties of a material. Planets are richer, more complex and more resistant to reductionist thinking. Genetics is the science of heredity and variation in biological systems. By analogy, we can speak of the genetics of a planet such as Earth, while also acknowledging that environment has a role in its evolution and its current state.

Cosmologists are familiar with thinking about time logarithmically: a lot happened in a very short period of time back near the Big Bang. To some extent, it helps to think about planet formation in a similar way (Fig. 1). The events that defined Earth's formation and the initial conditions for its subsequent evolution are squeezed into an epoch that may have already been over within 100 million years of the formation of the Solar System. In this epoch more happened inside Earth and more energy was dissipated from within the planet than throughout all of subsequent geological time. We have no direct geological record of this earliest epoch in the form of rocks and must rely instead on other sources of evidence.

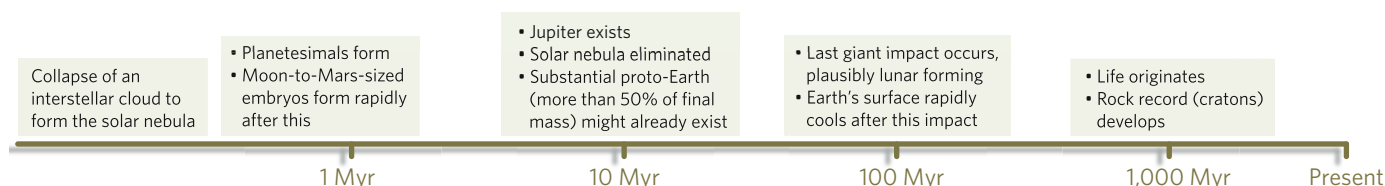
## Making planets

Our understanding of planet formation involves four major inputs: astronomical observations of places where planetary systems may currently be forming, the study of meteorites that formed even before the epoch in which the Solar System's planets formed, study of the planets themselves (Earth among them), and theoretical modelling. None of these is very complete or satisfying. The astronomical observations tell us about disks and dust and only indirectly about possible planets, the meteorites come from parent bodies that were probably always in orbits beyond Mars and are not necessarily representative of Earth's building blocks, Earth itself is good at concealing its history (through frequent surface rejuvenation), and theory is often either too permissive (many adjustable parameters) or falls short of a correct description of process. Even so, a picture emerges that has undergone considerable testing and refinement in recent years.

Current models of planetary formation<sup>2–6</sup> have had some success in explaining observations and have the following features. Almost 4.6 billion years ago, an interstellar cloud of gas and dust collapsed under the action of gravity. Angular momentum guaranteed that the collapse would be into a disk around the forming star (the Sun) rather than merely into the Sun alone. This disk had a radius of perhaps 50 astronomical units (AU), where 1 AU is the distance between Earth and the Sun. Almost all the mass of this disk was outwards of the eventual orbit of Earth. The particular mix of elements was nothing unusual, having been set by nucleosynthesis for the heavy elements and the outcome of the Big Bang for the lightest elements. Conversion of the gravitational energy of infall into heat assured that temperatures would be high in the inner part of the disk, sufficient to vaporize much of the infalling dust. Subsequent cooling allowed the formation of dust embedded in the primarily hydrogen gas. Through gentle collisions, these particles aggregated into larger particles up to a centimetre or more in size.

Meteoritic evidence strongly indicates the formation of larger 'planetesimals' that were kilometres or more in size, on a timescale of less than a million years. This process is poorly understood: planetesimals may have arisen through gentle collision and sticking of smaller grains or they may have arisen through gravitational instabilities in the disk. Such processes are presumed to have occurred throughout the Solar System. The timing of formation of these bodies is well established at around 4,567 million years ago, and their collapse from the interstellar medium can only have occurred a million years or less before this because of evidence for the presence of short-lived radioactive elements. This precisely determined date therefore well defines the origin of the Solar System.

Almost a billion bodies 10 km in diameter would be needed to make an Earth. However, it is not thought likely that planetesimals were the actual building blocks of Earth. A dense swarm of such bodies in nearly circular low-inclination orbits is gravitationally unstable on a short timescale. In less than a million years, much larger bodies ('planetary embryos') are formed that are Moon- to Mars-sized. These arise because of gravitational focusing of impacts between bodies with low relative velocity. Outward from the asteroid belt, these embryos may



**Figure 1 | A logarithmic view of the time of planetary formation.** The left end corresponds to the initiation of a collapse to form the solar nebula, and is close to 4,567 million years ago. Much happened in the 1 to 100 million

years (Myr) immediately after this, and the logarithmic scale correctly emphasizes the importance of this 100-Myr period, despite the shortness of this period compared with Earth's age.

exceed Earth in mass, but in the terrestrial zone they are still well short of Earth's final mass. This means we must build Earth from a modest number (100 or fewer) of these embryos, but adding in a sprinkling of planetesimals. The aggregation of embryos to even bigger bodies takes far longer than their formation, extending from tens of millions of years to as much as 100 million years, because it requires the excitation of eccentric orbits so that the embryos have an opportunity to collide<sup>3–5</sup>.

Earth and its companion terrestrial planets are a tiny part of the Solar System and it should come as no surprise that the presence of the giant planets, especially Jupiter, the most massive and closest of these to Earth, would have a role in Earth's formation. Jupiter must have formed while the hydrogen-dominated gas of the solar nebula was mostly still present<sup>6</sup>, and astronomical observations suggest that the gas may have been present in sufficient abundance for about 5 million years at most. We could perhaps imagine that the formation of Earth postdated the formation of Jupiter, and some models are of this kind. Realistically, a full understanding of Earth's formation probably requires a full understanding of Jupiter's formation. Jupiter is enriched in heavy elements relative to the Sun, and some part of that enrichment is likely to be present as a core. It is likely, although not certain, that this core was formed first, with the gas then placed on top. But whichever story is correct, the formation of Jupiter involves much more than the physics involved in building bodies such as Earth because we must understand gas accretion as well as the accretion of solids. At present, this understanding is incomplete. Models of Earth accretion are in many ways much more detailed than models of giant-planet formation, but they are contingent on understanding Jupiter.

### Planetary embryology

We have evidence about some of the planetesimals because they are the presumed source of most meteorites, but the much larger embryos have not left direct evidence of their existence. Nonetheless, it is likely that their properties are important for understanding Earth. They formed so quickly that they probably partly melted, owing to the presence of the short-lived radioactive isotope <sup>26</sup>Al. They may even have been big enough to undergo melting by the conversion of gravitational energy of formation into heat. Partial melting can be expected to cause the separation of a liquid iron alloy from the partly molten silicate mantle, and these embryos may even have had atmospheres. In short, they are planets with iron cores, short-lived but possessing properties derived from planetary processes rather than the properties of the precursor planetesimals. These differences from planetesimals can arise in a number of ways: ingassing (the incorporation of solar nebula gas, should the surface of the embryo be molten), the role of pressure (the mineral phases within the embryo and its crust can be different from those in a low-pressure planetesimal because of self-gravity), and the loss of material by escape (either because of high temperatures or through collisions). Close encounters, tidal disruption and the creation of debris during collisions are processes that are not currently well incorporated into models of planet formation.

The embryos responsible for forming Earth were not — indeed could not have been — built from planetesimals that formed at 1 AU, because the coalescence of the embryos necessarily requires their scattering around the inner part of the Solar System<sup>3–5</sup>. It is therefore incorrect to think of Earth's provenance and composition as being precisely defined,

and different from, say, those of Venus. On the other hand, some differences are expected purely by chance and, importantly, it is thought unlikely that any of the Earth-forming embryos formed out at locations where water ice could condense. Indeed, Earth is relatively dry, at least for the water inventory that we can measure (the oceans and upper mantle), and our water may have arisen through water-bearing planetesimals coming from greater distances rather than through water incorporated in the primary embryos. This remains somewhat controversial, and one of the goals of Earth science is to get a better understanding of Earth's complete water budget.

### Giant impacts and lunar formation

The likely dominance of the embryos as building-blocks for Earth implies the predominance of giant impacts. We should not think of Earth's formation as the steady accumulation of mass but rather as a series of infrequent, highly traumatic events separated by periods of cooling and healing. The largest, and possibly the last, of these events is thought to have been responsible for the formation of the Moon<sup>7,8</sup> (Fig. 2). Recent isotopic evidence<sup>9</sup> now dates this event at as much as 100 million years after the origin of the Solar System. Many features of the event would also apply to earlier non-lunar-forming events, except that those would have been less extreme. The impact origin of the Moon was once a controversial idea, but it has gradually been accepted for two reasons: the lack of a realistic alternative, and growing evidence for its compatibility with the data — isotopic data in particular. Particularly importantly, it is thought to set the stage for Earth's subsequent evolution.

The lunar-forming collision plausibly involved the oblique impact of a Mars-mass planetary embryo (10% of Earth's mass) with the ~90% complete Earth. The impact velocity would probably have been dominated by the infall into the mutual gravity field, and most of this energy would have been converted into heat. Unlike energy, angular momentum is much more nearly conserved throughout geological time, and this kind of impact explains well the current angular momentum of the Earth–Moon system. The mean temperature rise of Earth resulting from this collision can be estimated as  $\Delta T \approx 0.1 GM/RC_p \approx 4,000$  K, where  $G$  is the gravitational constant,  $M$  and  $R$  are Earth's mass and radius, respectively, and  $C_p$  is the specific heat of rock. Previous impacts would have heated Earth up to a hot, nearly isentropic state (a state in which entropy is nearly uniform with depth) close to, or partly in excess of, melting. Convective cooling below the freezing point is inefficient, so the state immediately before impact is hot, except perhaps right at the surface.

We expect that the impact heating would have been uneven because the various parts of Earth would be shocked to differing extents, but the immediate post-giant-impact state would relax to a very hot configuration, in which all or most of the rock and iron is in molten form and some silicate (perhaps even tens of per cent) is in vapour form. In most simulations of this kind of impact, a disk forms, derived mostly from the impacting body. For the expected radiating surface area and radiating temperature (~2,000 K), the cooling time to remove about half of the impact energy is around 1,000 years, perhaps somewhat shorter for the disk. This is a very short period relative to the time between major collisions, but a very important one. During this short period, the Moon forms, most of the core of the projectile merges with the core of the proto-Earth, some of the pre-existing Earth's atmosphere may be blown off, and a significant part of the deep, initially molten, mantle

of the Earth will freeze without having the opportunity to differentiate (because the crystals are advected vigorously by the turbulent convective motions that accompany the cooling).

The Moon probably did not form immediately after the giant impact, even though orbital times for material placed about Earth are less than a day. Instead, it seems to be necessary to wait for hundreds to thousands of years, the timescale of disk cooling, as it is thought likely that the Moon did form completely molten. For reasons not fully understood, the need to cool the disk is of greater importance than the shorter timescales of dynamical evolution. Perhaps lunar formation should not be thought of as disconnected from the provenance and evolution of the deep Earth. The reason is that, after the giant impact, some exchange of material may have taken place between Earth and the disk, aided by the vigorous convection of both the liquid and vapour parts of each and the presence of a common silicate atmosphere. This picture of rapid exchange makes the disk more Earth-like, rather than like the projectile that was responsible for its formation. The picture was originally motivated by a desire to understand the remarkable similarity of Earth and Moon oxygen isotopes<sup>8</sup> but also finds support in tungsten<sup>9</sup> and possibly silicon<sup>10</sup> isotopic evidence. However, we do not yet have a fully integrated model of lunar formation that is dynamically satisfactory as well as chemically acceptable.

### Core formation

The core-formation events (one event per giant impact) are particularly important because core formation is the biggest differentiation process of Earth: it involves one-third of Earth's mass and a large energy release, because the iron is about twice as dense as the silicates. To a substantial extent, it also defines the composition of Earth's mantle. In the immediate aftermath of a giant impact, we expect a substantial part of the core of the projectile to be emulsified with the molten mantle of the pre-impact proto-Earth. The core and mantle materials are thought to be immiscible (like water and oil) despite the very high temperatures, perhaps as high as 10,000 K for some of the material. If the material is mixed down to a small scale (perhaps even to the point where there are centimetre-sized droplets of iron immersed in the liquid silicate) then the iron and silicate can chemically and thermally equilibrate at high temperature and pressure (Fig. 3a). The composition of the core and the iron content of the mantle were presumably set during these equilibration episodes. The silicon and hydrogen contents of the mantle may also be affected by this equilibration, as both are soluble in iron at high pressure and temperature. These elements are particularly significant: silicon content affects the mineralogy of Earth's mantle, and the fate of hydrogen may have much to say about the total water inventory of Earth at this early epoch and the flow of mantle rocks. However, much of Earth's water may have been delivered later.

It is likely that some of the projectile iron is not mixed down to the smallest scales but instead finds its way to the core just hours after the impact (Fig. 3b). This iron will not equilibrate, either thermally or chemically, and it thus carries a memory of previous core-forming events at earlier times in smaller bodies (the embryos discussed earlier). The emerging picture is a complex one in which we should not expect the core or mantle of Earth to have a simple chemical relationship that involves the last equilibration at a particular pressure and temperature, but rather to have been formed under a range of thermodynamic conditions involving a number of significant events at different times<sup>2,11</sup>.

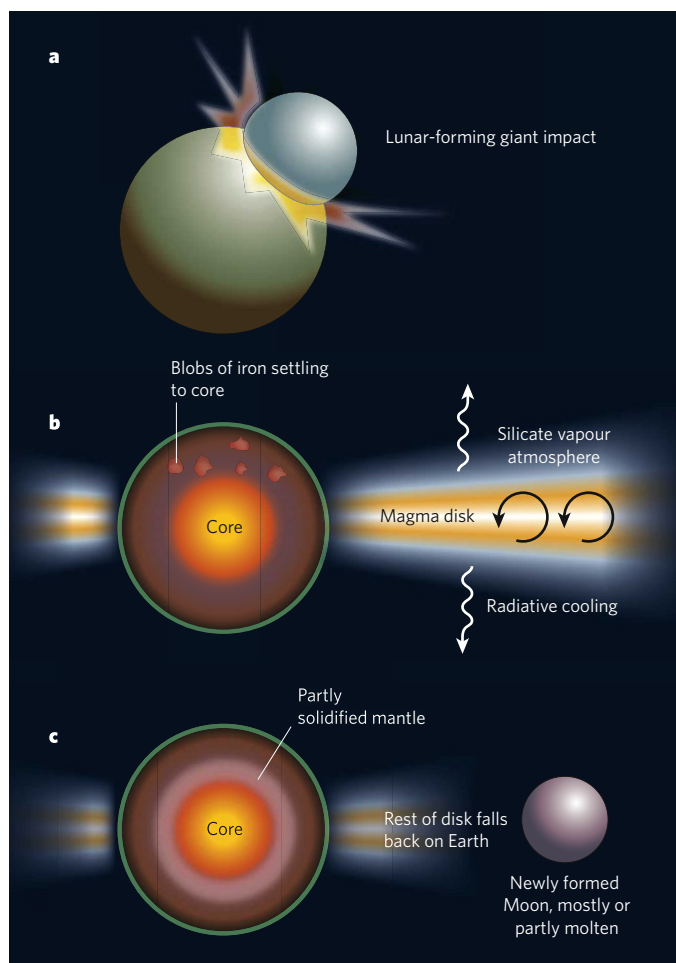
Earth's atmosphere at the time of a giant impact might have been mostly steam and carbon dioxide (CO<sub>2</sub>) — probably both were important. It is possible, but not certain, that a large part of the atmosphere was blown away immediately after the giant impact. Water vapour is, however, much more soluble than CO<sub>2</sub> in magma, so that even if the atmosphere were ejected into space, outgassing from the underlying magma ocean would replenish much of it. An important feature of water vapour is that it has a strong greenhouse effect, and that may have allowed the retention of an underlying magma ocean, even for the long periods between giant impacts. However, this type of atmosphere can rain out if there is insufficient energy supplied to its base (sunlight alone is insufficient) and,

as a consequence, any steam atmosphere may collapse on a geologically short timescale, leading to an Earth surface that is actually cool (able to have liquid water) even while the interior is very hot.

### Mantle differentiation

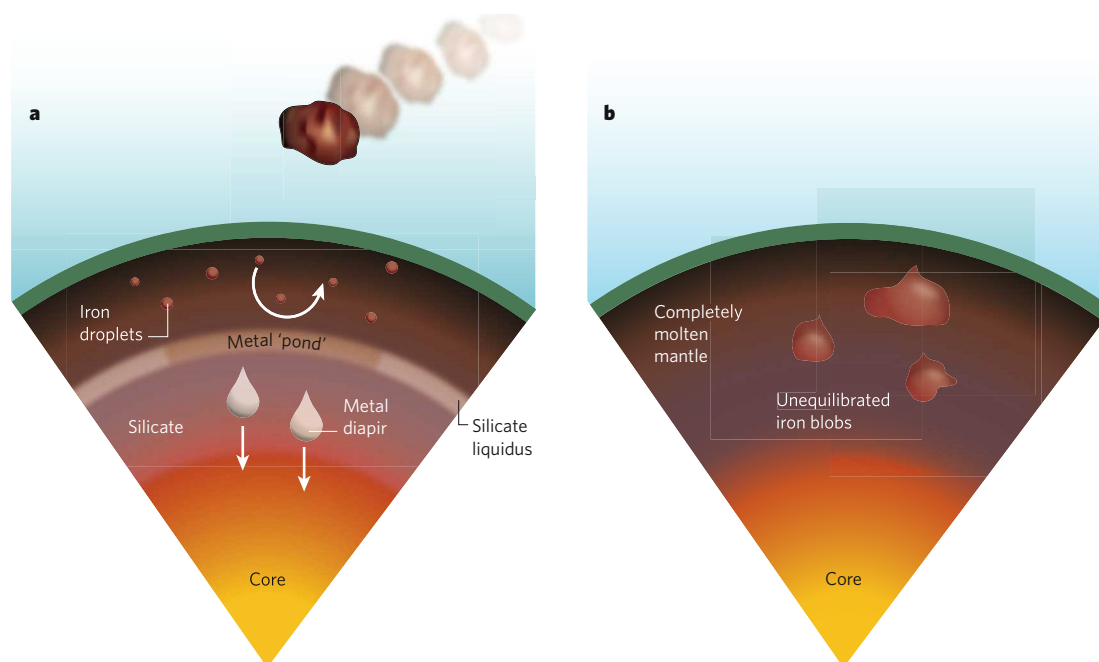
The mantle of the post-giant impact Earth will cool very fast at first<sup>12</sup>, limited only by the black-body radiation that can escape from the top of the transient (initially silicate vapour) atmosphere. The thermal structure of the mantle is expected to be close to isentropic because that is the state of neutral buoyancy and therefore the state preferred by convection, provided that viscosity is low. The nature of the freezing within this convecting state is of great importance and is thermodynamically determined. Many materials have the property that if they are squeezed isentropically, they undergo freezing even as they get hotter. Equivalently, they melt if they are decompressed isentropically from a frozen but hot, high-pressure state. The former correctly describes the freezing of Earth's solid inner core (the hottest place in Earth, yet frozen) whereas the latter correctly describes the melting responsible for the generation of basaltic magma, the dominant volcanism on Earth and most voluminously expressed at the low mantle pressures immediately beneath mid-ocean ridges. Recent work<sup>13,14</sup> suggests that this picture may not apply for the deeper part of Earth's mantle, so that freezing may begin at mid-depths.

Even so, there will eventually come a point (perhaps as soon as a few thousand years) after a giant impact when the bottom part of the mantle



**Figure 2 | The effect on Earth of the giant impact that formed the Moon.**

**a**, A giant planetary embryo collides with the nearly complete Earth. **b**, A magma disk is in orbit about Earth, while blobs of iron from the planetary embryo settle down through the mantle to join the existing core. **c**, The outermost part of the magma disk coalesces to form the Moon as the result of radioactive cooling, while the rest falls back to Earth. Inside Earth, the mantle nearest the core has partly solidified, and the mantle might acquire a layered structure.



**Figure 3 | Two contrasting views of what might have happened during core formation.** **a**, There is a magma ocean bounded below by a mostly solid lower region: the dispersed iron aggregates before descending to the core. **b**, Some of the iron from the core of the projectile responsible for a giant impact is imperfectly mixed and descends to the core on a short timescale as distorted blobs hundreds of kilometres in diameter, without equilibration with the mantle.

is mostly frozen. A very important question then arises: does the interstitial melt of this two-phase medium move up or move down under the action of gravity? It is very unlikely to be immobile. It is likely that it goes down (most probably because it is richer in iron than the coexisting solid), but in either case the mantle will differentiate internally into a layered structure (Fig. 4). This does not necessarily mean that Earth developed a primordial layering that has been preserved throughout geological time and is perhaps present still as part of the complex structure observed at the base of the mantle by seismologists and given by them the unromantic name of *D''* (see page 269). An early differentiation event for the silicate portion of Earth is favoured by some geochemists<sup>15</sup>, although, interestingly, it may have been earlier and it may have involved the formation of a primordial crust. It could perhaps be the cumulative consequence of giant impact events, a rare example of an Earth memory that even pre-dates the last giant impact.

The 'average' Earth surface environment during accretion may not have been very hot, even though there were undoubtedly short periods of time during which it was so hot that rocks were vaporized. These traumatic events reset the clock for subsequent evolution and emphasize the importance of the last such global event. Soon after the last global traumatic event, it may even be possible to have had rocks that survived throughout subsequent Earth history. Certainly, zircons — tiny, very resistant parts of rocks — have been dated back to ~4.4 billion years<sup>16</sup>, and it is not unreasonable to expect zircon discoveries that date back to within a few hundred million years of the lunar-forming impact. Zircons are not the same as hand specimens and rocks that can be studied in context (an intact structure, such as a surviving craton), but the gap is closing between the geological record as usually defined and the events that can only be dated through gross isotopic signals for Earth as a whole.

### Core memory

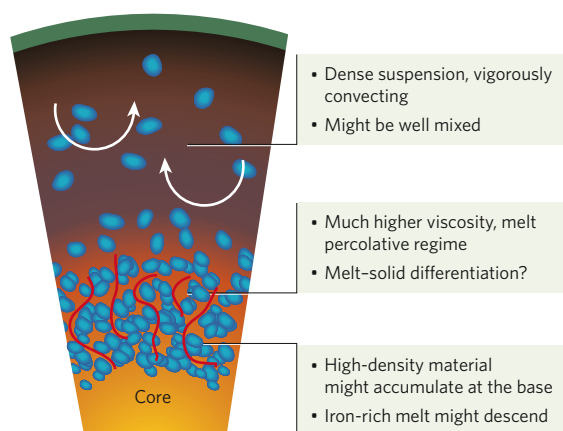
The composition of Earth's core is different from pure iron–nickel and this is presumably because of the modest solubility of other elements, especially oxygen, silicon and sulphur. The simplest view of Earth's core is that it is a hot fluid cooled from above. Significantly, Earth's core has superheat: it is hotter than the temperature it would have been if liquid iron alloy coexisting with upper to mid-mantle silicates had sunk isentropically to the core. We can estimate this superheat by knowing the temperature at which iron alloys freeze at the known pressure of Earth near its centre and by the seismological determination of the size of its inner core. This superheat is currently about 1,000 K or so, and

may initially have been in excess of 2,000 K. Unlike the mantle, the core cannot lose energy directly to the surface or to space and it is therefore likely that part of this superheat is a memory of the primordial Earth and may be telling us something about the specific processes responsible for core formation. Loss of primordial heat, together with the latent heat released as the inner core freezes, is potentially sufficient to maintain convection in the outer core over geological time, although even this is in some doubt given currently favoured values of the thermal conductivity of the core. In addition, buoyancy can be provided by the exclusion of part of the light elements from the inner core or perhaps from material exsolving from the outer core and attaching itself to the mantle.

Earth's magnetic field is generated by a dynamo: vigorous convection in the liquid, electrically conducting, outer core amplifies the existing magnetic field and thereby balances the tendency of the electrical current and associated fields to undergo decay. It is possible that these energy sources were insufficient to generate Earth's magnetic field even for the period when we know it must have existed<sup>17</sup>. A modest amount of radiogenic heat, most plausibly from the decay of  $^{40}\text{K}$ , is a suggested solution to the short-fall. The experimental support for this is equivocal, but given the possibly high temperatures for part of the core-forming materials, it may be more difficult to keep things out of the core (that is, to avoid the core becoming too low in density) than to get them in! The amount of potassium needed would be modest and so it might not be apparent as a marked depletion of potassium in Earth's mantle relative to elements of similar volatility. The ability of Earth to generate a magnetic field may also be linked to the presence of an efficient mechanism for eliminating heat through the planet's surface. Plate tectonics is a particularly efficient mechanism.

### Plate tectonics and life

We understand why Earth's mantle convects: there is no alternative mechanism for eliminating heat. However, we do not understand why Earth has plate tectonics. It is sometimes described as merely a property of the particular form that mantle convection takes on our planet, but this begs the question. Plate tectonics is neither mandatory nor common (there is no clear evidence of its existence on any other planet so far). Nonetheless, many think its presence is deterministic: given the specific parameters of present-day Earth, it is the behaviour expected, in the same sense that a physicist setting up a convection experiment on a layer of fluid heated from below need not be concerned about whether his chosen fluid was once a vapour or a solid. Even in this point of view, the presence of plate tectonics is history-dependent. For example, the amount and distribution of water may be important, as it is well established that water in rocks



**Figure 4 | Mantle cooling and differentiation during the later stages of a magma ocean.** As the magma of the mantle cools, a stage is eventually reached at which dense iron-rich interstitial liquid (red) percolates through the solid matrix (blue) to accumulate just above the core.

has a major effect on their melting properties and response to stress. Earth's water budget is likely to be dependent on its history. The surface environment is profoundly influenced by the presence or absence of a plate-tectonic cycle, and that environment is, in turn, influencing the existence of life and is then affected by the presence of life. Everything affects everything else: the development of life on Earth is not likely to be disconnected from the composition of Earth's core.

### Where do we go from here?

The remarkable advances over recent years and decades have been notable for their strongly interdisciplinary character, and some of this advance has come about through thinking of Earth as a planet and relating it to the environment in which it formed. Even so, the biggest challenge seems to require looking inside the planet: we need to understand better the phase relationships between Earth's constituents, the way in which mantle convection works and how to integrate this with plate tectonics, the connection between the deep Earth and our ocean and atmosphere, and the generation of Earth's magnetic field. The origin

and development of life are also clearly questions for Earth science and will resist compelling answers until we have better characterized the thermodynamic, chemical and fluid dynamical environments. The deep Earth is deeply significant and also deeply informative for Earth's surface and all of Earth science. ■

David J. Stevenson is in the Division of Geological and Planetary Science, California Institute of Technology, Pasadena, California 91125, USA.

1. Udry, S. *et al.* The HARPS search for southern extra-solar planets — XI. Super-Earths (5 and 8  $M_{\oplus}$ ) in a 3-planet system. *Astron. Astrophys.* **469**, L43–L47 (2007).
2. Halliday, A. N. & Wood, B. J. in *Treatise on Geophysics* Vol. 9 (ed. Schubert, G.) 13–50 (Elsevier, Amsterdam, 2007).
3. Chambers, J. E. Planetary accretion in the inner Solar System. *Earth Planet. Sci. Lett.* **223**, 241–252 (2004).
4. Raymond, S. N., Mandell, A. M. & Sigurdsson, S. Exotic Earths: Forming habitable worlds with giant planet migration. *Science* **313**, 1413–1416 (2006).
5. Ogiwara, M., Ida, S. & Morbidelli, A. Accretion of terrestrial planets from oligarchs in a turbulent disk. *Icarus* **188**, 522–534 (2007).
6. Lissauer, J. J. & Stevenson, D. J. in *Protostars and Planets V* (eds Reipurth, B., Jewitt, D. & Keil, K.) 591–606 (Univ. Arizona Press, Tucson, 2007).
7. Canup, R. M. Dynamics of lunar formation. *Annu. Rev. Astron. Astrophys.* **42**, 441–475 (2004).
8. Pahlevan, K. & Stevenson, D. J. Equilibration in the aftermath of the lunar-forming giant impact. *Earth Planet. Sci. Lett.* **262**, 438–449 (2007).
9. Touboul, M., Kleine, T., Bourdon, B., Palme, H. & Wieler, R. Late formation and prolonged differentiation of the Moon inferred from W isotopes in lunar metals. *Nature* **450**, 1201–1209 (2007).
10. Georg, R. B., Halliday, A. N., Schauble, E. A. & Reynolds, B. C. Silicon in the Earth's core. *Nature* **447**, 1102–1106 (2007).
11. Rubie, D. C., Nimmo, F. & Melosh, H. J. in *Treatise on Geophysics* Vol. 9 (ed. Schubert, G.) 51–90 (Elsevier, Amsterdam, 2007).
12. Solomatov, V. in *Treatise on Geophysics* Vol. 9 (ed. Schubert, G.) 91–119 (Elsevier, Amsterdam, 2007).
13. Stixrude, L. & Karki, B. Structure and freezing of  $MgSiO_3$  liquid in Earth's lower mantle. *Science* **310**, 297–299 (2005).
14. Mosenfelder, J. L., Asimow, P. D. & Ahrens, T. J. Thermodynamic properties of  $Mg_2SiO_4$  liquid at ultra-high pressures from shock measurements to 200 GPa on forsterite and wadsleyite. *J. Geophys. Res.* **112**, B06208 (2007).
15. Boyet, M. & Carlson, R. W. Nd-142 evidence for early (> 4.53 Ga) global differentiation of the silicate Earth. *Science* **309**, 576–581 (2005).
16. Wilde, S. A., Valley, J. W., Peck, W. H. & Graham, C. M. Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Gyr ago. *Nature* **409**, 175–178 (2001).
17. Nimmo, F. in *Treatise on Geophysics* Vol. 9 (ed. Schubert, G.) 217–241 (Elsevier, Amsterdam, 2007).

**Author Information** Reprints and permissions information is available at [npg.nature.com/reprints](http://npg.nature.com/reprints). Correspondence should be addressed to the author ([djs@gps.caltech.edu](mailto:djs@gps.caltech.edu)).

# Using seismic waves to image Earth's internal structure

Barbara Romanowicz

**Seismic waves generated in Earth's interior provide images that help us to better understand the pattern of mantle convection that drives plate motions.**

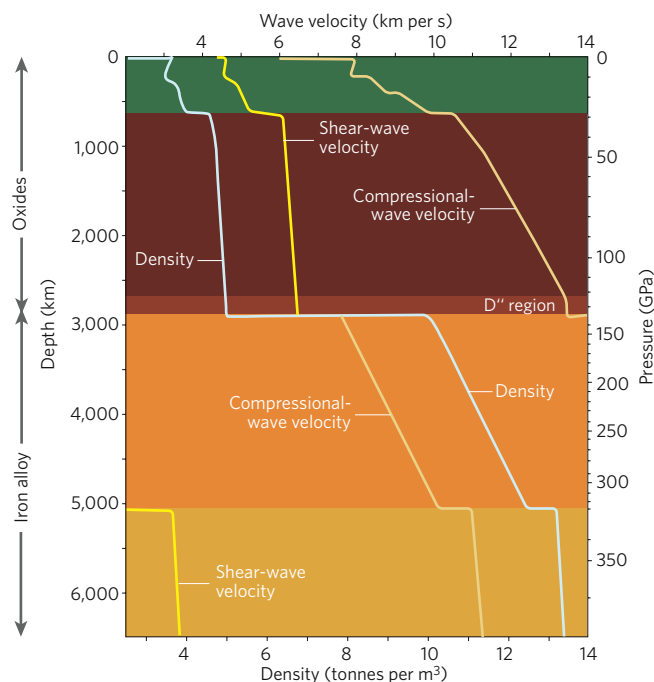
Forty years after the discovery of seafloor spreading and the acceptance of the theory of plate tectonics, important gaps remain in our understanding of the pattern of convection that drives the motions of the plates, leading to earthquakes, tsunamis and volcanic eruptions. There are still many heated debates. Does oceanic lithosphere pushed down into the interior at converging plate boundaries reach the bottom of Earth's mantle? Do deep-rooted, thin hot plumes rise through the mantle under mid-plate 'hot spot' volcanoes? What is the relative importance of compositional versus thermal heterogeneity in mantle convection? And what role does Earth's solid inner core have in the 'geodynamo', which keeps Earth's magnetic field alive, and in the thermal evolution of our planet (see page 261)? To address these controversies, seismology has been brought to bear to image Earth's deep interior. From the construction of accurate models of Earth's one-dimensional radial structure (Fig. 1) to the current models of its three-dimensional structure (Fig. 2), progress in seismic imaging has gone hand in hand with improvements in the design of seismic sensors, the capacity to record digitally increasingly massive quantities of data, theoretical progress in handling seismic-wave propagation through complex three-dimensional media and the development of powerful computers for simulating seismic waves and for the inversion of large matrices.

From seismic tomography, first introduced in the late 1970s (refs 1, 2), we now have a good understanding of the first-order characteristics of the long-wavelength (~1,000–2,000 km) three-dimensional elastic structure of Earth's mantle<sup>3</sup>. At shorter wavelengths (~200 km), fast-velocity 'slabs' representing oceanic lithosphere plunging back into the mantle are, today, the best-resolved 'objects', because of the favourable geometry; many earthquake sources illuminate such slabs from both below and above, at least down to ~600 km depth (Fig. 3a). It is tempting to interpret the large-scale features imaged throughout the mantle in terms of lateral variations in temperature, which can be as much as several hundred degrees Celsius. For example, the fast ring of high velocities at the bottom of the mantle (shown in blue in Fig. 2) might well represent the 'graveyard' of cold subducted lithosphere, and the slow regions, commonly referred to as 'superplumes', the hot rising return flow (shown in red in Fig. 2). It is increasingly clear, however, that compositional variations also have an important role in mantle convection.

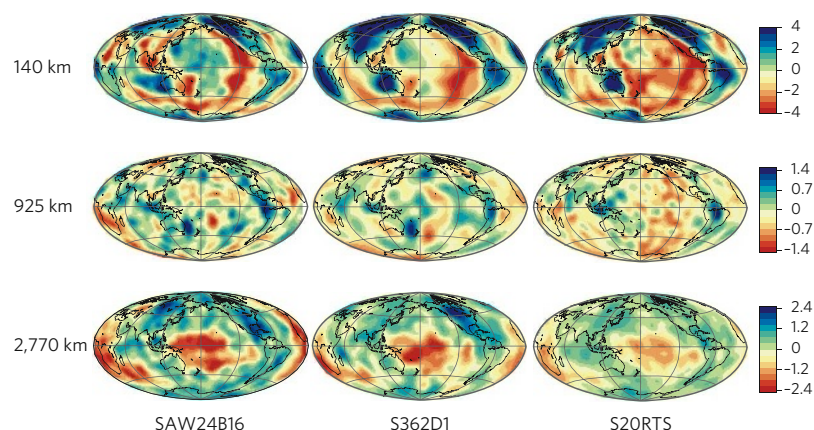
With the deployment, starting in the early 1980s, of high-quality digital broadband seismic stations around the world (Fig. 3b), finer-scale imaging became possible. Particularly striking is the accumulating evidence for complexity in the lower 300–400 km of the mantle, the so-called D'' region, an important chemical and thermal boundary layer. Many intriguing seismic observations have been made in this region<sup>4,5</sup>, including the remarkable observation that the lateral transition from fast shear velocity regions in D'' into the superplumes occurs abruptly, over a much smaller range than would be possible if lateral

variations in temperature were the only cause<sup>6</sup>. Perhaps less surprisingly, closer to Earth's surface such strong lateral contrasts are also found at lithospheric depths, especially at the edges of tectonic provinces of different origin and age.

Characterizing the sharpness or fuzziness of the boundaries of the heterogeneous structures deep inside the planet, and detecting and mapping small-scale heterogeneity, are the next steps. This will mean extracting more information from seismograms than has traditionally been done. Indeed, neither remnants of compositionally distinct lithosphere in the lower mantle nor narrow plume conduits (if they exist) can be accurately mapped by standard tomographic approaches that make use only of information carried by the most direct waves — those that travel along the shortest paths — according to the simple



**Figure 1 | Radial structure of Earth.** The first-order structural units of Earth — its suite of concentric shells and their approximate composition — were established over the first half of the twentieth century from measurements of the travel times of seismic waves refracted and reflected inside Earth, whereas proof of the solidity of the inner core had to await the capability to record and digitize long time series and measure the frequencies of free oscillations. The '660 km' discontinuity is a phase change, and possibly a compositional change, in the silicate mantle. This illustration is of the preliminary reference Earth model<sup>14</sup>.



**Figure 2 | Large-scale three-dimensional Earth structure as inferred from seismic tomography.** Each column of images represents a different model of mantle shear-velocity structure (using various data sources) shown at three representative depths (140 km, 925 km and 2,770 km). Left, SAW24B16, developed at the University of California at Berkeley<sup>15</sup>; centre, S362D1, developed at Harvard University<sup>16</sup>; and right, S20RTS, developed as a collaboration between the University of Oxford and California Institute of Technology<sup>17</sup>. In the top ~250 km of the mantle, the structure follows the surface tectonics: slow ridges and back-arcs (red), fast roots under stable continents (blue) and a progressively faster velocity away from mid-ocean ridges, consistent with expectations from a simple cooling plate model. Below the thickest lithospheric roots (250 km), the pattern changes, and in the transition zone a clear signature of fast anomalies associated with subducted slabs emerges. Recent models show a variety of behaviours for these slabs: some seem to be stagnating in the upper mantle; for others, the fast-velocity anomaly seems to continue at oblique or steep angles into the lower mantle. Two regions, in northwestern America and Southeast Asia, show fast-velocity anomalies that may be related to past subduction down to considerable depths (~1,200–1,400 km). In the mid-mantle, the spectrum of heterogeneity becomes white, which indicates that it is dominated by smaller-scale features. In the bottom 1,000 km of the mantle, as we approach the core–mantle boundary, a new pattern of long-wavelength heterogeneity progressively emerges, with two very large antipodal low-velocity regions centred in the Pacific Ocean and under Africa and surrounded by faster than average material. The units of the key are relative shear-velocity changes (as percentages) with respect to the global mean at the given depth.

rules of ray theory. It will be necessary to take account of the energy bouncing off weak scatterers that can have a wide range of sizes. In practice, this means working in a wide frequency band, at short spatial wavelengths, using both the amplitude and the travel times of all possible seismic phases — that is, the entire seismogram — and applying signal-enhancing techniques.

A significant challenge is the limited distribution of seismic-wave sources and receivers. Ideally, one would want to sample the volume of Earth uniformly. But unlike other disciplines that use imaging, such as medical tomography or petroleum exploration, earthquake seismologists cannot optimize their experimental geometry (Fig. 3). To overcome these limitations, several promising approaches are being pursued.

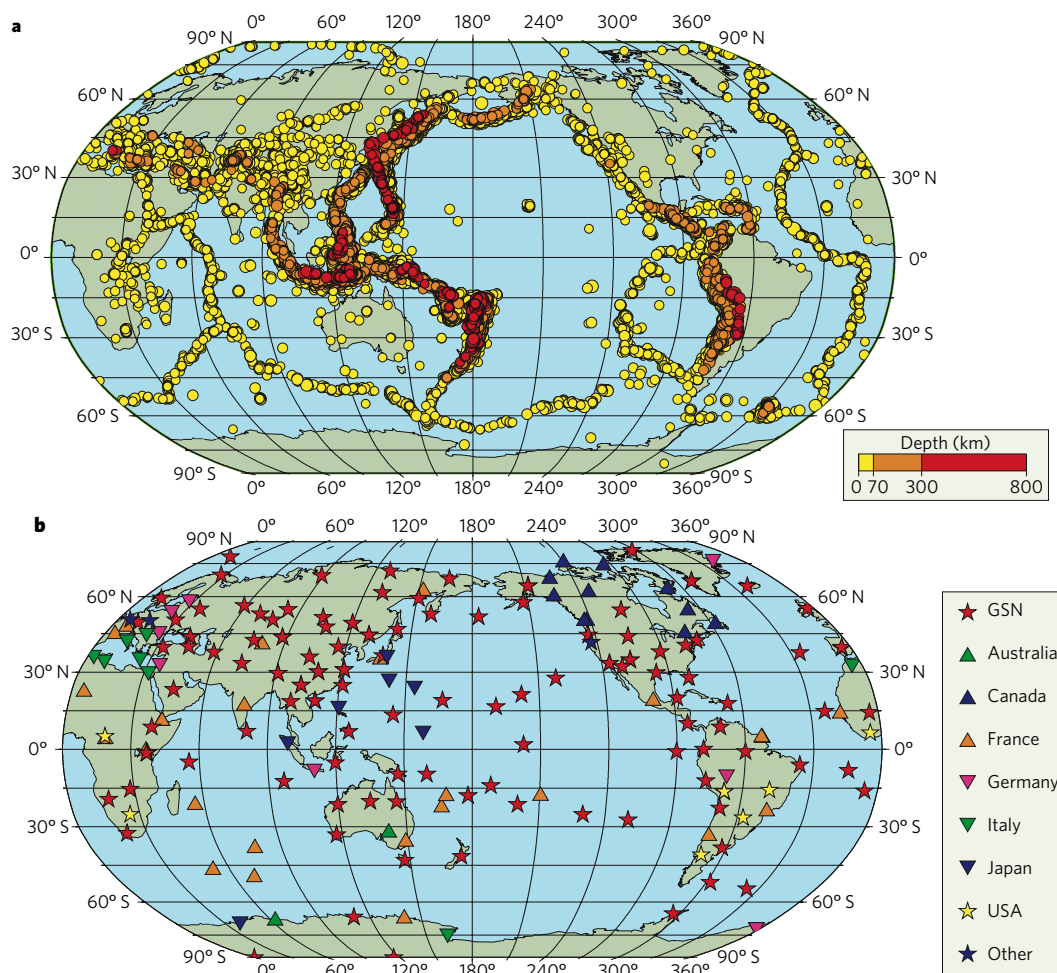
New and exciting horizons have recently opened up with increasing capabilities in both computation and data collection. There are now powerful numerical schemes to compute synthetic seismograms in structures of arbitrary complexity, such as the spectral element method<sup>7</sup>, which are well adapted to the spherical global geometry of Earth. They can be used in a variety of ways, for forward modelling of observed seismic waveforms, as well as for inversion of the seismogram to retrieve the three-dimensional structure. They are still heavy on computation but hold much promise for the construction of the next generation of global tomographic models. Anisotropy and dissipation, which also influence seismic-wave propagation, can now be better characterized and provide additional information on flow directions, temperature variations and the presence of partial melting. At the higher end of the seismic spectrum, the deployments of dense permanent regional arrays, such as Hi-net in Japan, or temporary ones such as those of PASSCAL (<http://www.iris.edu/about/PASSCAL>), are stimulating the

development of techniques that are beginning to erode the difference between global seismology and exploration geophysics.

Through the utilization of energy scattered both backward and forward, impressively detailed images of slabs are starting to be constructed. For the first time, it is possible to use the results of seismic imaging to trace the fate of water as it is entrained down into the mantle with the subducting slab<sup>8</sup>. The global seismic network, complemented by PASSCAL-type deployments<sup>9,10</sup>, and local dense arrays provide sufficient spatial sampling in some continental areas to investigate fine-scale layering in the deep mantle using newly developed sophisticated back-projection techniques. Much is expected from the data set now being assembled through the USArray programme of Earthscope (<http://www.iris.edu/USArray>). Seismologists can start to put precise constraints on velocity contrasts and the sizes and depths of heterogeneous bodies. These can be combined with experimental and theoretical data about mineral physics to determine lateral variations in composition and temperature. For example, in the case of the recently discovered post-perovskite transition, which is thought to occur in the temperature/pressure range of the D'' region (see page 269), mineral physicists and geodynamicists are working hand in hand with seismologists to search for its presence in the deep mantle and evaluate its consequences for mantle dynamics<sup>4,5</sup>.

The approaches mentioned above assume that appropriately distributed earthquake sources are available. Where this is not possible, a rapidly developing technique to eliminate the constraints associated with natural earthquakes is building on the data set of continuous broadband waveforms accumulated by many stations in the world. Background seismic noise continuously excited by the oceans and the atmosphere can be used to construct tomographic images through noise cross-correlation. The promise of this approach has been demonstrated in the investigation of the crust<sup>11</sup>, for which the presence of strong energy in the microseismic frequency band (~1–15 s) can be exploited. A possible extension of the technique to longer-period seismic waves presents interesting prospects for imaging the upper mantle at high resolution down to at least the base of the lithosphere.

This still leaves the oceans, where recording is limited to sparsely distributed islands. Yet there are key geodynamic problems to be addressed: for instance, the deep structure and anisotropy of ocean basins are not well understood. Most volcanic hot spots are in the oceans. The recent controversy about the 'banana-doughnut kernel' technique<sup>12</sup> indicates the level of frustration: improvements in wave-propagation theory and inclusion of scattering effects cannot make up for the fact that stations on hot-spot islands are isolated, so that it is not possible to accurately constrain the depth and lateral extent of underlying slow anomalies. Many areas in the deep mantle and the core are currently not accessible because of a lack of stations in the oceans. Although efforts to instrument the ocean floor have been ongoing for more than 20 years, long-term ocean-floor broadband stations are still few. Local temporary deployments, such as those beneath mid-ocean ridges, have led to spectacular results<sup>13</sup>, and other ongoing projects, such as the Plume project in Hawaii, will help to address specific targets. A cabled observatory is planned in the northwest Pacific, combining Canadian and US efforts (<http://www.orionprogram.org/OOI/default.html>). But an internationally coordinated programme is needed to systematically deploy large-aperture (1,000 km × 1,000 km) broadband ocean-floor arrays that would be left in place for at least one or two years, to record a sufficient number and variety of earthquakes and progressively fill the gap in illuminating deep structure under the oceans.



**Figure 3 | Global seismicity and networks.** **a**, Worldwide distribution of earthquakes of magnitude ( $M_w$ ) greater than 5.0 from 1 January 1991 to 31 December 1996. Earthquakes occur mainly along plate boundaries, delineating, in particular, the global mid-ocean ridge system. Earthquakes are generally shallow (yellow). In subduction zones around the Pacific Ocean and in the collision zones in southern Eurasia, intermediate-depth (orange) and deep (red) earthquakes indicate the presence of cold lithospheric slabs

plunging into Earth's mantle. **b**, The current global broadband digital seismic network (shown as at October 2007) has been constructed through an international effort coordinated by the Federation of Digital Seismic Networks (FDSN), complemented by denser permanent regional arrays (not shown) and temporary regional deployments. GSN, Global Seismic Network (the US component of the international network). (Panel **b** courtesy of R. Butler, IRIS, Washington DC.)

Finally, as the images provided by seismologists become sharper, there is an increasing opportunity to work closely with other geoscientists — geochemists, geodynamicists and mineral physicists — to make the best of complementary constraints for the challenging 'inverse problem' that the interior of our planet represents — that is, to use observations at or near the surface of Earth to constrain ideas about its deep structure and dynamics. Better communication and cross-education among these disciplines is key to progress. This is why interdisciplinary programmes such as the Cooperative Institute for Deep Earth Research (<http://www.deep-earth.org>) are needed. ■

Barbara Romanowicz is at the Berkeley Seismological Laboratory and the Department of Earth and Planetary Science, University of California at Berkeley, 215 McCone Hall, Berkeley, California 94720, USA.

1. Dziewonski, A. M., Hager, B. H. & O'Connell, R. J. Large scale heterogeneities in the lower mantle. *J. Geophys. Res.* **82**, 239–255 (1977).
2. Aki, K., Christofferson, A. & Husebye, E. Determination of the three-dimensional structure of the lithosphere. *J. Geophys. Res.* **82**, 277–296 (1977).
3. Romanowicz, B. Global mantle tomography: progress status in the last 10 years. *Annu. Rev. Geophys. Space Phys.* **31**, 303–328 (2003).
4. Lay, T. *et al.* The core mantle boundary layer and deep mantle dynamics. *Nature* **392**, 461–468 (1998).
5. Hirose, K. Postperovskite phase transition and its geophysical implications. *Rev. Geophys.* **44**, RG3001, 18p (2006).

6. Wen, L. Seismic evidence for a rapidly varying compositional anomaly at the base of the Earth's mantle beneath the Indian Ocean. *Earth Planet. Sci. Lett.* **194**, 83–95 (2001).
7. Komatitsch, D., Ritsema, J. & Tromp, J. The spectral element method, Beowulf computing and global seismology. *Science* **298**, 1737–1742 (2002).
8. Kawakatsu, H. & Watada, S. Seismic evidence for deep-water transportation in the mantle. *Science* **316**, 1468–1471 (2007).
9. Bostock, M. G. *et al.* An inverted continental Moho and serpentinization of the forearc mantle. *Nature* **417**, 536–538 (2007).
10. Van der Hilst, R. *et al.* Seismostratigraphy and thermal structure of Earth's core-mantle boundary region. *Science* **315**, 1813–1817 (2007).
11. Shapiro, N. *et al.* High resolution surface-wave tomography from ambient seismic noise. *Science* **307**, 1615–1618 (2005).
12. Kerr, R. A. Rising plumes in Earth's mantle: phantom or real? *Science* **313**, 1726 (2006).
13. Forsyth, D. W., Webb, S. C., Dorman, L. M. & Shen, Y. Phase velocities of Rayleigh waves in the MELT experiment on the East Pacific Rise. *Science* **280**, 1235–1238 (1998).
14. Dziewonski, A. M. & Anderson, D. L. Preliminary reference Earth model. *Phys. Earth Planet. Inter.* **25**, 297–356 (1981).
15. Mégnin, C. & Romanowicz, B. The 3D shear velocity structure of the mantle from the inversion of body surface and higher mode waveforms. *Geophys. J. Int.* **143**, 709–729 (2000).
16. Gu, Y. J. A., Dziewonski, M., Su, W.-J. & Ekström, G. Models of the mantle shear velocity and discontinuities in the pattern of lateral heterogeneities. *J. Geophys. Res.* **106**, 11169–11199 (2001).
17. Ritsema, J., van Heijst, H. J. & Woodhouse, J. H. Complex shear wave velocity structure imaged beneath Africa and Iceland. *Science* **286**, 1925–1928 (1999).

**Author Information** Reprints and permissions information is available at [npg.nature.com/reprints](http://npg.nature.com/reprints). Correspondence should be addressed to the author ([barbara.romanowicz@gmail.com](mailto:barbara.romanowicz@gmail.com)).

# Mineralogy at the extremes

Thomas S. Duffy

**The discovery of a new silicate structure at conditions corresponding to a depth of 2,700 kilometres below Earth's surface has fundamentally changed our understanding of the boundary between the core and mantle.**

Connections between scientific disciplines can emerge in unexpected ways. In 2004, mineralogists rushed to their libraries to locate a somewhat obscure 40-year-old paper<sup>1</sup> that described an unusual crystal structure found in a compound of calcium iridium oxide ( $\text{CaIrO}_3$ ). The reason for the sudden geological interest in the iridate family was the discovery that  $(\text{Mg,Fe})\text{SiO}_3$  perovskite — the major mineral in Earth's vast lower mantle — adopted this same structure when subjected to pressures of more than 125 GPa (1.25 million bars) and temperatures above 2,000 K in the laboratory<sup>2,3</sup>. Under these crushing pressures and searing temperatures, Earth's mantle finally divulged one of its deepest secrets. The new structure, commonly referred to as post-perovskite, is composed of layers of  $\text{SiO}_6$  octahedra sharing edges and corners to form sheets interleaved with layers of larger Mg and Fe cations (Fig. 1). Although mineralogists had speculated over the years that perovskite might undergo some kind of transformation at high pressures, the formation of this  $\text{CaIrO}_3$ -type structure had been wholly unanticipated by theory and experiment.

## New view of the deep Earth

Earth's lower mantle, which extends from a depth of 660 km to 2,890 km, is the largest region of Earth, with a mass that is roughly 100 times that of the crust. Understanding the mineralogical constituents of this region is vital to unravelling Earth's origin, evolution and dynamic behaviour (see page 261). Without any way to sample it directly, our fuzzy picture of the lower mantle comes mainly from seismic studies, and most of the region seems to be fairly homogeneous. However, a puzzling aspect has been a thin layer extending about 200 km above the boundary between the core and the mantle (known for historical reasons as D'') that has several anomalous properties<sup>4</sup>. The D'' region is separated from the rest of the mantle by a discontinuity in seismic velocity. Compared with the rest of the lower mantle, the D'' region is very heterogeneous and has increased anisotropy of seismic waves (see page 266). Complexity in the deepest mantle should not be surprising. The hot but solid silicate minerals of the mantle are juxtaposed against the churning liquid iron core. The region is a likely source for the hot plumes that reach all the way to Earth's surface, as well as perhaps the final repository for subducting slabs from Earth's surface.

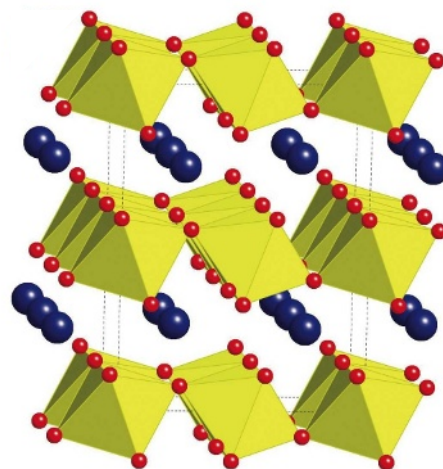
So what has been learned about the connection between D'' and post-perovskite in the three years since its discovery? On balance, many of post-perovskite's characteristics match those predicted by seismic observations of D'' (ref. 5). Although it is difficult to measure pressure accurately under such extreme conditions in the laboratory, the transformation seems to occur at pressures corresponding to those found at the top of the D'' region. More importantly, the strongly positive pressure-temperature, or Clapeyron, slope of the transition means that the transformation occurs deeper in locally hotter regions and shallower in cooler regions, which is consistent with seismic observations. But it can be much more complex than this. Earth has a steep thermal gradient near the core-mantle boundary, and temperatures at the base of the mantle might become hot enough for perovskite to re-emerge just above the core<sup>6</sup>. In this case, complex structures such as localized lenses of post-perovskite could be expected

(Fig. 2). Attempts to image the structures in this region seismically have already yielded some tantalizing results<sup>7,8</sup>.

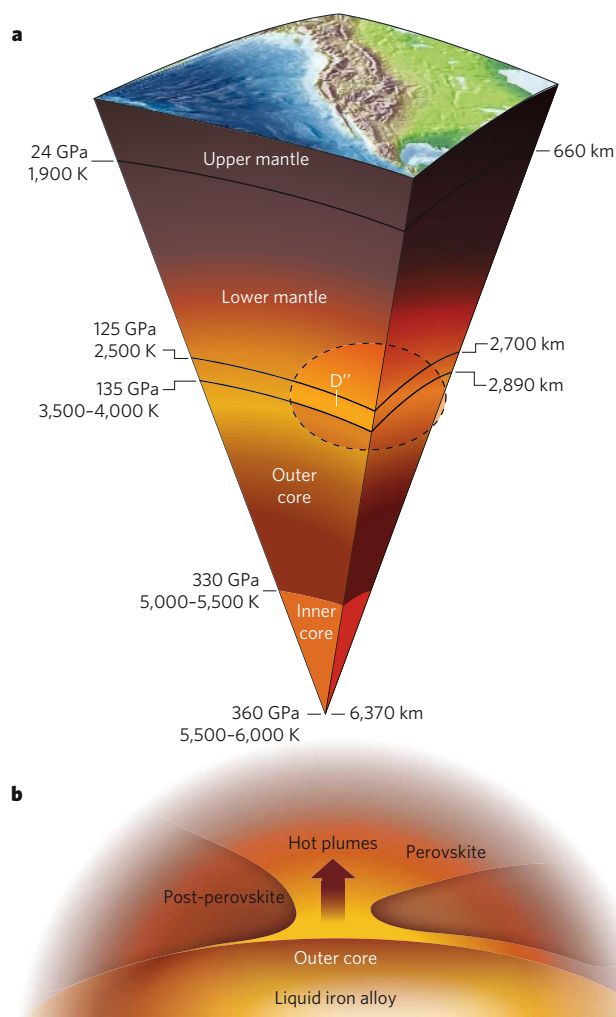
## Going beyond the core-mantle boundary

Are there more discoveries of the magnitude of post-perovskite awaiting us in the deep Earth? The answer to this question is almost certainly yes. Several trends are fuelling a vibrant and vigorous research enterprise in the exploration of deep planetary interiors and the wider high-pressure realm. In the laboratory, sustained pressures in excess of 1 Mbar (relevant to Earth's deep mantle and core) can be achieved with a diamond anvil cell. However, mineralogists are now finding that they can carry out increasingly reliable studies under extreme conditions without experimental input, by using computer calculations based on quantum-mechanical principles, such as density-functional theory<sup>9</sup>. The major advantage of such methods is that they can simulate pressure conditions of 1 Mbar nearly as easily as they can simulate 1 bar. The disadvantage is that the theory's inherent approximations mean that the results have to be compared with experiments. Theoretical studies have provided tremendous insights into post-perovskite — confirming its thermodynamic stability and providing predictions of the Clapeyron slopes, seismic anisotropies and other key properties, some of which have yet to be confirmed experimentally. Rapid improvements in theoretical methods and their applications to increasingly complex systems will certainly be a major driving force for the field in the coming years.

In laboratory studies carried out at high pressures, the megabar era has now been entered. Pressures above 1 Mbar (100 GPa), which until recently were the domain of a determined few, are now just the starting point for much forefront science. Pressures in Earth's interior range up to 360 GPa, and temperatures are perhaps 5,500–6,000 K near Earth's centre. Much of the deep mantle and core thus remain *terra incognita* from



**Figure 1 | Crystal structure of the post-perovskite phase of  $(\text{Mg,Fe})\text{SiO}_3$ .** The structure consists of layers of linked silicon octahedra (yellow). Red spheres at vertices of  $\text{SiO}_6$  octahedra are oxygen ions, and blue spheres are magnesium and iron ions.



**Figure 2 | Cross-section through Earth's interior showing the expected range of pressures and temperatures. a,** The lower mantle extends from a depth of 660 km to a depth of 2,890 km, with the D'' region extending about 200 km above the core. **b,** A simplified diagram of possible structures of the D'' region near the core-mantle boundary (the region indicated by dashed lines in **a**)<sup>6</sup>.

an experimental perspective. To progress, a coupled effort is required to achieve and sustain a well-characterized pressure-temperature state while making sophisticated measurements of a range of key physical variables on both solid and liquid phases, including structure, elasticity, bonding, transport properties, lattice dynamics, electrical and magnetic properties and chemical interactions among increasingly complex geological assemblages.

Key questions about Earth's core (which has a pressure range of 135–360 GPa) include the identity of its main light elements, the nature of melting and iron-rich liquids at core conditions, core-mantle interactions and the origin of the solid inner core's seismic anisotropy. Moreover, Earth cannot be studied in isolation. The interior structures of the giant planets present a myriad of fascinating questions, and their study requires even higher pressures and temperatures. For giant planets, the materials of main interest are the fundamental ices and gases (for example, hydrogen, water and methane) of the Solar System. Complexity abounds in these constituents, and new bonding configurations, structural changes and metallization are all expected<sup>10</sup>. Such studies can provide the answers to basic questions about the mechanisms of planetary formation and the origin of magnetic fields. Even further, new possibilities can be envisaged for the structures of hot 'Jupiters' and possible super 'Earths' and super 'Ganymedes' in solar systems beyond that of Earth, offering combinations of composition, pressures and temperatures that hold the promise of further surprises.

## Scaling up

Aside from the scientific opportunities, a key driving force for mineral physics has been the union of high-pressure experiments with synchrotron X-ray facilities<sup>11</sup>. High-pressure studies are especially well positioned to benefit from the combination of high-energy and high-intensity radiation that synchrotrons specialize in delivering. X-ray spectroscopy techniques that have matured at synchrotrons have found important applications in the Earth sciences. The discovery that iron in mantle minerals transforms from a high-spin (or unpaired) state to a low-spin (or paired) state is another finding of great importance<sup>12</sup>. The change in spin state is accompanied by changes in partitioning behaviour, compressibility and optical properties, all of which can strongly affect the behaviour of the lower mantle. This is a reminder that mineral properties can change markedly under extreme conditions even without any accompanying changes in crystal structure.

Synchrotrons are now focal points around which communities of high-pressure scientists nucleate. The result has been a flowering of interdisciplinary interactions. This trend towards community facilities promises to grow as new opportunities abound to bring high-pressure mineral physics to neutron facilities such as the Spallation Neutron Source at Oak Ridge, Tennessee, and laser facilities such as the National Ignition Facility in Livermore, California. It is worth emphasizing that static techniques are only one method of achieving ultra-high pressure-temperature conditions. Historically, high pressures were first reached by shock-wave methods that sustain extreme conditions for no longer than a microsecond. Dynamic methods are also undergoing a renaissance driven by new capabilities in high-powered lasers. These techniques are achieving multi-megabar conditions, and there is potential to reach much greater pressures by using these methods alone or together with diamond anvil technologies<sup>13</sup>.

The discovery of post-perovskite is likely to be remembered as a turning point in understanding the structure and dynamics of the deep Earth. But the elucidation of the connections between the geophysics of the deep Earth and its mineralogical constituents has only just begun. Given the fundamental questions that remain to be addressed, the unexplored territory of pressure-temperature-composition space and newly emerging scientific capabilities, post-perovskite promises to be just the first of many scientific highlights that will characterize the megabar realm of deep planetary interiors.

Thomas S. Duffy is in the Department of Geosciences, Princeton University, Princeton, New Jersey 08544, USA.

- Rodi, F. & Babel, D. Ternary Oxide der Übergangsmetalle 4. Erdalkali-iridium(4)-oxide Kristallstruktur von CaIrO<sub>4</sub>. *Z. Anorg. Allg. Chem.* **336**, 17–23 (1965).
- Murakami, M., Hirose, K., Kawamura, K., Sata, N. & Ohishi, Y. Post-perovskite phase transition in MgSiO<sub>3</sub>. *Science* **304**, 855–858 (2004).
- Oganov, A. R. & Ono, S. Theoretical and experimental evidence for a post-perovskite phase of MgSiO<sub>3</sub> in Earth's D'' layer. *Nature* **430**, 445–448 (2004).
- Garnero, E. Heterogeneity of the lowermost mantle. *Annu. Rev. Earth Planet. Sci.* **28**, 509–537 (2000).
- Wookey, J., Stackhouse, S., Kendall, J.-M., Brodholt, J. & Price, G. D. Efficacy of the post-perovskite phase as an explanation for lowermost-mantle seismic properties. *Nature* **438**, 1004–1007 (2005).
- Hernlund, J. W., Thomas, C. & Tackley, P. J. A doubling of the post-perovskite phase boundary and structure of the Earth's lowermost mantle. *Nature* **434**, 882–886 (2005).
- Lay, T., Hernlund, J., Garnero, E. J., Thorne, M. S. A post-perovskite lens and D'' heat flux beneath the central Pacific. *Science* **314**, 1272–1276 (2006).
- van der Hilst, R. D. et al. Seismo-stratigraphy and thermal structure of Earth's core-mantle boundary region. *Science* **315**, 1813–1817 (2007).
- Oganov, A. R. et al. *Ab initio* theory for planetary materials. *Z. Kristallogr.* **220**, 531–548 (2005).
- Scandolo, S. & Jeanloz, R. The centers of planets. *Am. Scientist* **91**, 516–525 (2003).
- Duffy, T. S. Synchrotron facilities and study of Earth's deep interior. *Rep. Prog. Phys.* **68**, 1811–1859 (2005).
- Badro, J. et al. Iron partitioning in Earth's mantle — toward a deeper lower mantle discontinuity. *Science* **300**, 789–791 (2003).
- Jeanloz, R. et al. Achieving high-density states through shock-wave loading of precompressed samples. *Proc. Natl Acad. Sci. USA* **104**, 9172–9177 (2007).

**Acknowledgements** G. Shen (Carnegie Institution of Washington) and S.-H. Shim (Massachusetts Institute of Technology) provided helpful comments.

**Author Information** Reprints and permissions information is available at [npg.nature.com/reprints](http://npg.nature.com/reprints). Correspondence should be addressed to the author ([duffy@princeton.edu](mailto:duffy@princeton.edu)).

# Earthquake physics and real-time seismology

Hiroo Kanamori

**The past few decades have witnessed significant progress in our understanding of the physics and complexity of earthquakes. This has implications for hazard mitigation.**

Simply stated, an earthquake is caused by slip on a fault. However, the slip motion is complex, reflecting the variation in basic physics that governs fault motion in different tectonic environments. Seismologists can learn a great deal about earthquakes from studying the details of slip motion.

## The size of great earthquakes

Seismic slip motion involves a broad 'period' (or frequency) range, at least from 0.1 s to 1 hour, and a wide range of amplitudes, roughly from 1  $\mu$ m to 30 m. Most seismographs available before the 1960s could record ground motions over only short periods — less than 30 s — which prevented seismologists from studying important details of earthquake processes. In the 1960s, longer-period analogue seismographs became available, allowing seismologists to study great earthquakes (in general, magnitude  $\geq 8$ ) over an extended period range; this has resulted in substantial progress in our understanding of earthquakes. For example, being able to measure long-period waves of up to 1 hour has made it possible for seismologists to establish the overall size of great earthquakes accurately. With the old instruments, wave amplitudes were measured over only a short period range, leading to underestimates of the magnitude of great earthquakes (Fig. 1). The monitoring of long-period waves, combined with rigorous use of wave theory, rectified this problem, and, as a result, our perception of global seismicity changed drastically during the twentieth century. With the old estimates, global seismic activity seemed to have been relatively constant over the century, but according to the new estimates, a burst of activity occurred between 1952 and 1965; about 40% of the total seismic energy released during the century was released during this period (see ref. 1 for a review).

Another important finding is that most earthquakes involve a relatively low stress change, 1–10 MPa. This contrasts with the much higher stress — 100 MPa or greater — involved in fracture of rocks at high confining pressure. This indicates that the fracture process in Earth's crust involves special physics, and this is currently the subject of extensive research.

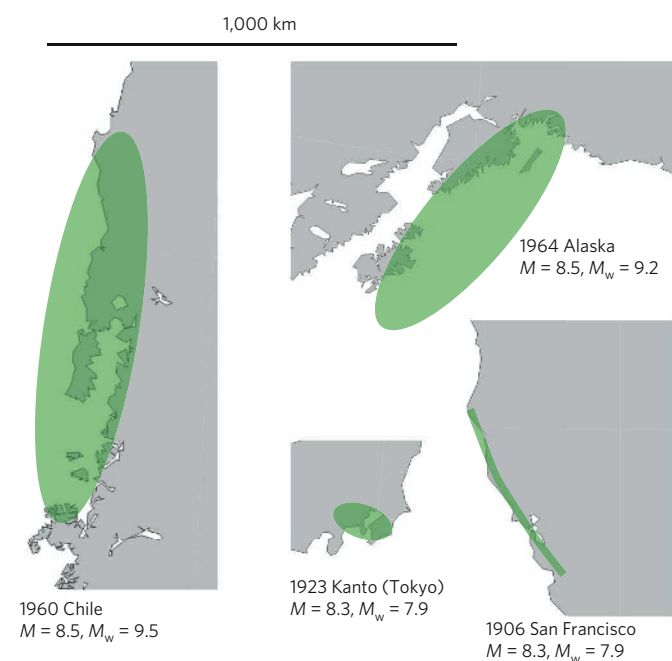
A better understanding of great earthquakes has also contributed to an improved understanding of the relationship between earthquakes and global plate motion.

## Earthquake diversity

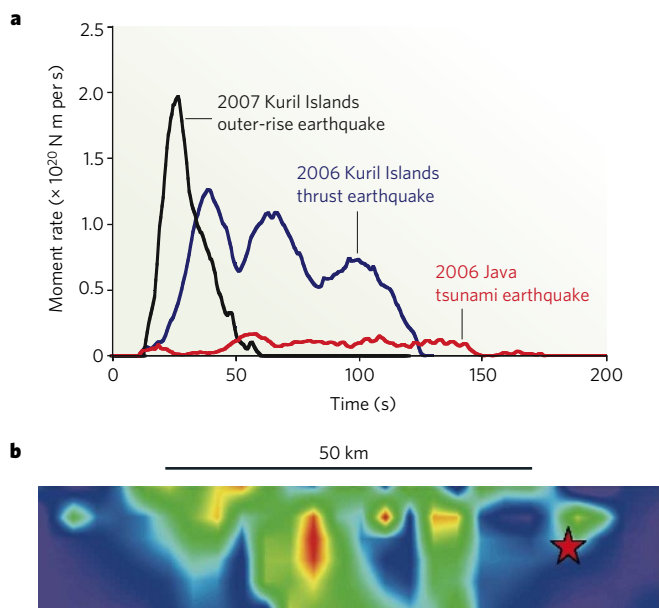
Since the late 1970s, force-balanced seismographs, which can record ground motions over a very broad period range (from 0.02 s to hours) and have a large dynamic range in amplitude (a factor of  $10^7$  in the ratio of the smallest to the largest amplitude)<sup>2</sup>, have become widely used. These instruments revolutionized observational seismology, and the fact that they are networked on a global scale, as well as a regional scale, has been especially useful (see page 266). Studies with broadband seismographs revealed a remarkable diversity of earthquakes in terms of slip characteristics and energy budget. Figure 2a shows the temporal diversity of energy release in earthquakes. Some earthquakes slip slowly and others rapidly, depending on the tectonic environment in which they occur. This diversity not only has important implications

for seismic hazard but also provides important clues about the fundamental physics of earthquakes. Subduction-zone earthquakes with slow slip tend to generate unexpectedly large tsunamis (for instance, the 2006 Java tsunami, represented by a red curve in Fig. 2a). Those earthquakes that occur within the subducting slab (for example, the Kuril Islands earthquake of 2007, denoted by a black curve in Fig. 2a) tend to have faster slip and can cause much stronger shaking than those with comparable magnitudes that occur on the subduction boundary (for example, the Kuril Islands earthquake of 2006, denoted by a blue curve in Fig. 2a). At present, these special characteristics are not fully considered in hazard-mitigation practices, and they need to be more explicitly considered in the future.

With an increased density of strong-motion seismographs deployed in many seismically active areas, together with geodetic data obtained using global positioning systems (GPS) and satellite interferometry,



**Figure 1 | Comparison of rupture area and magnitude.** Use of very long-period ( $> 200$  s) waves removed the saturation problem of the old surface-wave magnitude,  $M$ , which saturates above 8 and gives approximately the same values for great earthquakes regardless of the size of the rupture area inferred mainly from the aftershock area.  $M_w$  is the magnitude determined using long-period waves with a rigorous source theory and is known as the moment magnitude. Illustrated in this figure are the comparisons of the rupture areas (green) and the magnitude (both  $M$  and  $M_w$ ) for the 1960 Chilean earthquake, the 1964 Alaskan earthquake, the 1923 Kanto (Tokyo) earthquake and the 1906 San Francisco earthquake. The aftershock region of the 2004 Sumatra–Andaman earthquake ( $M_w = 9.2$ ) was about 1,400 km long — even longer than that of the 1960 Chilean earthquake.



**Figure 2 | Temporal and spatial diversity of seismic slip.** **a**, Moment rate function, which is roughly proportional to the energy release rate at the source as a function of time, for three earthquakes. The blue curve represents a great (magnitude ( $M_w$ ) 8.3) subduction-zone thrust earthquake that occurred in the Kuril Islands on 15 November 2006. This behaviour is typical of most earthquakes that occur on a boundary between an oceanic and a continental plate. The black curve is for a great ( $M_w = 8.1$ ) outer-rise earthquake that took place in the Kuril Islands on 13 January 2007. This event occurred within the subducting plate. The red curve represents a slow tsunami earthquake that occurred off the coast of Java on 17 July 2006. (Panel modified, with permission, from ref. 13.) **b**, Spatial rupture pattern of the Landers earthquake that hit California on 28 June 1992 (ref. 14). The star indicates where the rupture began. Red areas indicate the patches with the largest slip, about 6 m. The spatial variation of slip indicates the variation of stress and frictional properties, which can be used to study rupture physics. These properties also control the strength and frequency content of strong ground motions. N m, newton metre.

more details of seismic slip motion became clear. In old fault models, fault slip was treated as a spatially uniform slip propagating at a constant speed. This picture is still useful for understanding the general properties of fault motion, but it is very simplistic. The slip heterogeneity revealed by recent studies is often characterized by terms such as 'asperities' and 'barriers', as shown in Fig. 2b, which demonstrates the spatial diversity of seismic slip. Asperities are the portions on a fault at which large slip occurs, and barriers are patches where fault motion is impeded. Asperities and barriers reflect the heterogeneities of stress, the frictional properties of faults and geometries, and have a key role in the nucleation, growth and cessation of slip motion. The frictional properties depend on not only the static condition on the fault but also the slip velocity itself, and the resulting slip motion can exhibit highly complex patterns. Fault friction is the subject of active theoretical, laboratory and field studies, and various elementary processes including melting, fluid pressurization, fault lubrication and microfracturing have been examined<sup>3</sup>.

Sudden changes in rupture propagation caused by asperities and barriers control the strength and complexity of strong ground motions. Modern ground-motion estimates take advantage of the detailed slip models obtained recently. Models that take asperities and barriers into account are expected to provide much more realistic information about ground motion, which will be useful to engineers designing earthquake-resistant structures.

These studies using broadband seismographs also demonstrate the importance of taking into account the effects of long-period waves excited by large (in general, magnitude  $< 8$ ) and great (magnitude  $\geq 8$ ) earthquakes when designing tall buildings and large structures<sup>4,5</sup>.

### Towards multi-scale science

Most seismologically measured parameters are macroscopic in that they represent the quantity integrated over the entire fault motion. Such parameters include seismic moment (a quantity proportional to the product of the amount of slip and the fault area), radiated energy, fault dimension and the change in stress (that is, the stress drop). In cases in which near-field measurements are available, local slip functions and stress changes at every point on the fault can also be determined<sup>6</sup>. The relationship between the slip and the stress is called the 'fault constitutive relation' and can bridge the macroscopic fault parameters and the microscopic properties studied in theoretical, laboratory and field investigations. In this sense, seismology has become an intellectually challenging, multi-scale science that attempts to integrate traditional macroscopic seismological properties, medium-scale fault constitutive relations, and microscopic theoretical-laboratory-field parameters to obtain a comprehensive physical model of seismic rupture processes (see ref. 7 for more details).

### Slow and silent earthquakes

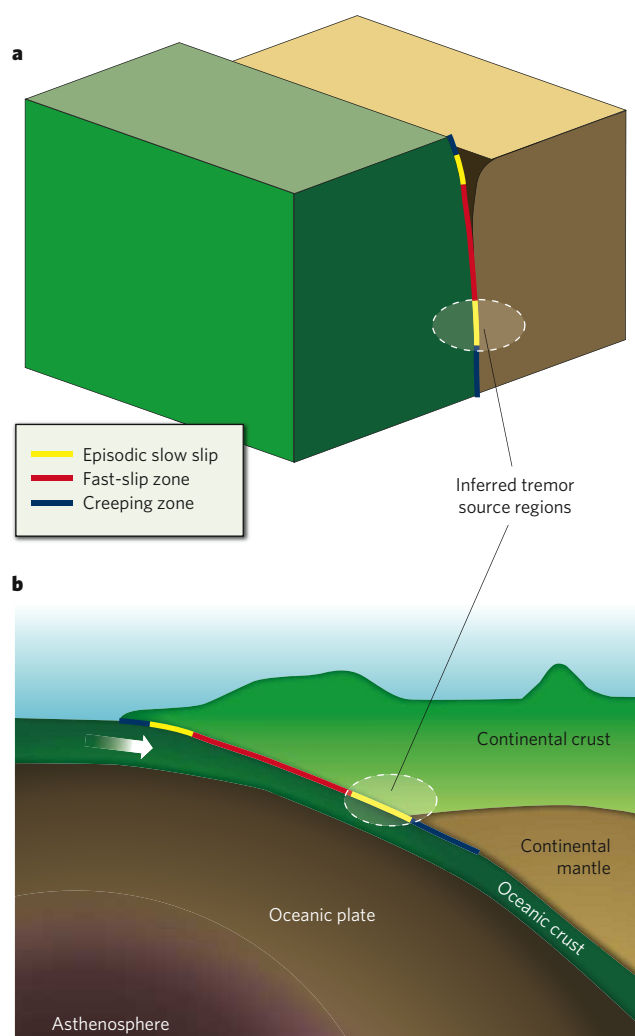
Recent studies using GPS and high-density seismic networks extended the measurable period range to days, months and years, which led to the discovery of slow and silent earthquakes<sup>8</sup>. From early seismological studies, some earthquakes were known to be slow, with a timescale longer than a few minutes, but these recent studies demonstrated the existence of seismic events with even longer timescales, which are often associated with small tremors<sup>9</sup>. It is generally agreed that these events occur on the downward extension of the seismogenic megathrust boundary at subduction zones and of crustal faults (Fig. 3). They represent the transitional behaviour from shallow brittle failure to deeper creeping motion. Many studies suggest that fluids released from hydrous minerals carried by the subducting slab are responsible for silent events at subduction zones. Although the details are still under extensive investigation, these events probably influence the state of stress in the adjacent seismogenic boundary, and the current interest is focused on whether the activity of silent events can provide a clue to the occurrence of large megathrust earthquakes in the same region. This problem is particularly important in the Cascadia subduction zone, which stretches, just offshore, along the northwest coast of North America, and the Nankai trough, off the coast of southwest Japan, where historical great megathrust earthquakes have been documented in detail, and where similar great earthquakes are certain to occur again.

Whether the physics of silent earthquakes is similar to, or entirely different from, that of regular earthquakes is an interesting scientific question. Although their timescales differ markedly, it is possible that the basic physics is the same and that the timescale difference is just a result of different energy partition between the radiated and dissipated energy. If so, what is responsible for the difference in partition? By contrast, the deformation mechanism can differ substantially: for example, brittle failure on a plane (regular earthquakes) versus volumetric slow deformation (silent earthquakes). This is one of the most exciting research topics at present.

### Real-time seismology and earthquake early warning

Even though seismologists have made considerable progress in understanding the basic physics of earthquakes, precise short-term earthquake predictions are still difficult to make because a large number of interacting elements are involved in the nucleation, growth and termination of an earthquake. At present, it is almost impossible to determine all the minute details that contribute to the occurrence of an earthquake. However, a better understanding of the overall physics of stress accumulation and release processes will improve our ability to carry out long-term forecasting of seismicity in many active areas in the world. With the accumulation of more data and improved methodology, long-term forecasting, in conjunction with improved engineering practice, will hopefully contribute significantly to the mitigation of seismic hazard in the future.

Significant progress has been made in the area of real-time seismology.



**Figure 3 | Locations of brittle fast-slip (seismogenic) and slow-slip zones. a, Vertical strike-slip fault. b, Subduction-zone boundary. (Figure adapted, with permission, from ref. 15).**

Real-time seismology refers to a practice by which we rapidly estimate, immediately after a significant earthquake, the source parameters and the distribution of shaking intensity, and distribute the information to various users. These users include emergency services officials, utility companies, transportation services, the media and the general public. This information will be useful for reducing the impact of a damaging earthquake on our society<sup>10</sup>.

In most cases, it takes minutes to hours to process the data, and by the time the information reaches the users, the damage may already have occurred at the user site. In this case, the information is called

post-earthquake information. This information is important for orderly recovery operations in the damaged areas. A good example of this post-earthquake information is ShakeMap (<http://earthquake.usgs.gov/eqcenter/shakemap>)<sup>11</sup>.

By contrast, if the data processing and information transfer can be done very rapidly (for instance, within 10 s), the information reaches some sites before shaking starts there. In such cases, the information is called 'earthquake early warning' (see ref. 12 for more details). This concept has been around for more than 100 years but was not put into practice until recently owing to technical and practical difficulties. In Japan, a warning system for impending ground shaking after a nearby large earthquake was implemented in the 1960s in conjunction with the operation of the high-speed bullet train. This system led to the subsequent development of earthquake early-warning methods for more general purposes. Several earthquake early-warning methods have been developed recently in many countries, and some have already been implemented. For example, in Japan, early-warning information is being distributed to the public and holds the promise of being a practical way to mitigate earthquake damage. Now that technical capability has been demonstrated, the next important step is to educate the public and to use early-warning information effectively through the judicious use of modern technology, such as control engineering.

Hiroo Kanamori is at the Seismological Laboratory, California Institute of Technology, Pasadena, California 91125, USA.

1. Kanamori, H. The diversity of the physics of earthquakes. *Proc. Jpn Acad. B* **80**, 297–316 (2004).
2. Wieland, E. & Streckeisen, G. The leaf-spring seismometer: design and performance. *Bull. Seismol. Soc. Am.* **72**, 2349–2367 (1982).
3. Rice, J. R. & Cocco, M. in *Tectonic Faults: Agents of Change on a Dynamic Earth* (ed. Handy, M. R., Hirth, G. & Hovius, N.) 446 (Massachusetts Inst. Technology Press, Cambridge, Massachusetts, 2007).
4. Olsen, K. B., Archuleta, R. J. & Matarrese, J. R. Three-dimensional simulation of a magnitude 7.75 earthquake on the San Andreas fault. *Science* **270**, 1628–1632 (1995).
5. Heaton, T. H., Hall, J. H., Wald, D. J. & Halling, M. W. Response of high-rise and base-isolated buildings to a hypothetical  $M_w$  7.0 blind thrust earthquake. *Science* **267**, 206–211 (1995).
6. Ide, S. & Takeo, M. Determination of constitutive relations of fault slip based on seismic wave analysis. *J. Geophys. Res.* **102**, 27379–27392 (1997).
7. Abercrombie, R., McGarr, A., Di Toro, G. & Kanamori, H. (eds) *Earthquakes: Radiated Energy and the Physics of Faulting: Geophysical Monograph 170* (American Geophysical Union, Washington DC, 2007).
8. Dragert, H., Wang, K. & James, T. S. A silent slip event on the deeper Cascadia subduction interface. *Science* **292**, 1525–1528 (2001).
9. Obara, K. Nonvolcanic deep tremor associated with subduction in southwest Japan. *Science* **296**, 1629–1681 (2002).
10. Kanamori, H. Real-time seismology and earthquake damage mitigation. *Annu. Rev. Earth Planet. Sci.* **33**, 195–214 (2005).
11. Wald, D. J. et al. TriNet 'ShakeMaps': rapid generation of peak ground motion and intensity maps for earthquakes in southern California. *Earthquake Spectra* **15**, 537–555 (1999).
12. Gasparini, P., Manfredi, G. & Zschau, J. (eds) *Earthquake Early Warning Systems* (Springer, Berlin, 2007).
13. Ammon, C. J., Kanamori, H. & Lay, T. A great earthquake doublet and seismic stress transfer cycle in the central Kuril islands. *Nature* doi:10.1038/nature06521 (in the press).
14. Wald, D. J. & Heaton, T. H. Spatial and temporal distribution of slip for the 1992 Landers, California, earthquake. *Bull. Seismol. Soc. Am.* **84**, 668–691 (1994).
15. Schwartz, S. Y. in *Treatise of Geophysics* (ed. Schubert, G.) (Elsevier, Oxford, in the press).

**Author Information** Reprints and permissions information is available at [npg.nature.com/reprints](http://npg.nature.com/reprints). Correspondence should be addressed to the author ([hiroo@gps.caltech.edu](mailto:hiroo@gps.caltech.edu)).

# From landscapes into geological history

Philip A. Allen

**Erosional and depositional landscapes are linked by the sediment-routing system. Observations over a wide range of timescales might show how these landscapes are translated into the narrative of geological history.**

Earth's landscape, shaped by the interplay between tectonics and climate, is a dynamic interface over which many biogeochemical cycles operate. The mass fluxes associated with the physical, biological and chemical processes acting across the landscape involve the transport of particulate sediment and solutes. Sediment is moved from source to sink — from the erosional engine of mountainous regions to its eventual deposition — by the sediment-routing system. The selective long-term preservation of elements of the sediment-routing system to produce the narrative of the geological record is dictated by processes operating in Earth's lithosphere. Making the connection between these two levels of enquiry — between the forces shaping present-day erosional and depositional landscapes and the long-term historical record — requires integration and ingenuity. If successful, we may indeed “see a world in a grain of sand” as the poet William Blake suggested.

The growing field of study of Earth surface processes is uniting the normally disparate disciplines of solid Earth geology, geomorphology and atmospheric and oceanographic sciences. Conference sessions are packed with contributions on Earth surface processes and new journal sections are devoted to it. These developments are not a result of a sudden conversion to an environmentalist agenda, but of a growing realization of the myriad of interactions, and the strength of the associated mass fluxes, that operate across the critical zone comprising Earth's surface. Understanding Earth surface processes therefore provides vital insights into how Earth functions as a system.

For Earth surface processes to be a vibrant new discipline, rather than a rebranding of conventional reductionist thinking, integration is required at different levels. One level is the integration of the physical, chemical and biological processes that shape Earth's surface and that drive its mass fluxes, investigated at the so-called human timescale — over the period for which we have historical records. The second level of integration is over larger spatial and temporal scales. Making the connections between these two levels is the exciting challenge that faces a wide range of natural scientists today.

## Sediment on the move

Earth's surface is the critical interface across which the bulk of Earth's chemical and biological exchanges take place. Most biogeochemical cycles involve the transport of material by fluids either in dissolved form or as particles of sediment. In fact, more than 20 billion tonnes of particulate sediment is delivered to the ocean every year, representing an average rate of loss of 135 tonnes per annum from each square kilometre of Earth's surface<sup>1</sup>. About the same mass is washed into the oceans by rivers as solutes every year, thereby controlling the ocean's bulk geochemistry, nutrient loading and biological productivity. Ignoring anthropogenic effects, this annual delivery to the coastal ocean is controlled by variations in global topography, climate and rock type, which are ultimately dependent on plate tectonics.

In the deep geological past, there were periods of pronounced convergence and collision of tectonic plates, when they organized themselves into great supercontinents before splitting and dispersing. One such period of assembly about 680–530 million years ago produced the great

supercontinent known as Gondwanaland. Erosion of the tectonic edifice caused by this amalgamation, which stretched for 10,000 km and was 1,000 km wide, resulted in the deposition of immense amounts of sand on the adjacent continental margins and in the deep sea<sup>2</sup>. The eroded material was equivalent to the blanketing of all of North America with a layer of sediment 10 km thick — an extraordinary image of the vigour of mass fluxes associated with Earth surface processes.

## The erosional engine of mountain landscapes

Sites of plate collision are typified by high mountains that act as the engine for mass fluxes of sediment over Earth's surface. The topography of mountains forms in the face of a relentless attack by erosion, which carves deep valleys into tectonically uplifting bedrock. Because erosion depends strongly on climatic factors, a goal of geoscience has been to decipher the distinctive imprint of climatic variations on mountainous landscapes. Intuitively, this seems straightforward enough. But it is not, as the problem is complicated by an issue that bedevils (or enriches, depending on your standpoint) a great deal of the discipline of Earth surface processes — the different temporal resolution of two interacting sets of processes: in this case tectonic fluxes and climatically driven erosion. A challenge for the future is to make progress in discovering the governing equations for erosion and resolving their time dependence.

The tectonics of mountain belts acts like a juggernaut: changes in tectonic conditions, such as in the direction or speed of relative motion of two colliding plates, are transferred very slowly to the deforming zone between them and to its surface topography. The time for the mountain belt to adjust to the new tectonic conditions may be several millions of years<sup>3</sup>. Climate, by contrast, is changeable and fickle. By the time the topographic surface has noticed that there are stirrings deep in the mountain belt, the climate may have changed along the roller-coaster of cold to warm, wet to arid many times over, as is well known from the approximately 100,000-year climatic periodicity of the late Pleistocene epoch. As a result, it is difficult to estimate the long-term release of particulate sediment from the erosional engine of mountains.

## Sediment-routing systems from source to sink

One can imagine tracking the trajectory of a single grain of sand from its source in mountain headwaters to its sink in a river flood plain, delta or the deep sea (Fig. 1). Each grain would have a different trajectory and a different time in transit. The integration of a multiplicity of such trajectories defines a sediment-routing system<sup>4</sup>, and an integration of the different transit times, were it possible, would provide information on the ability of the routing system to buffer incoming sediment flux signals. In other words, the sediment flux signal from the contributing upland river catchment is likely to be transformed, phase-shifted and lagged by the internal dynamics of the routing system. If this is the case, how can we possibly decipher the forcing mechanisms for a particular record in deposited sediment without knowing how it has been transformed by the internal dynamics of the sediment-routing system? We are right to be suspicious of oversimplistic interpretation of the ‘structure’ found in the large number of time-series records that geology throws up — for

instance, the mass accumulation rate of land-derived sediment in the ocean, or the packaging of genetic units of sedimentary rock. If the buffering timescale is greater than a million years for large river systems<sup>5,6</sup>, incoming sediment flux signals might be unrecognizable by the time they are propagated into the ocean.

Ideally, we would know all of the physical and chemical processes governing the sediment-routing system. This would be enormously gratifying in trying to understand how sediment-routing systems function generically, but we would immediately run into a fundamental problem: the long result of time. Time transforms sediment-routing systems into geology, and like history, selectively samples from the events that actually happened to create a narrative of what is recorded. Progress in understanding modern sediment-routing systems now leaves us poised to answer the important question: how do we simultaneously use the modern to generate the time-integrated ancient, and 'invert' the ancient to reveal the forcing mechanisms for change in the past?

### A world in a grain of sand

A growth area in the Earth sciences is the tracking of sediment from source to sink. We naturally ask, when picking up a handful of beach sand, 'where does this come from'? This question of origins is the science of provenance.

In the past, provenance analysis was centred on the general mineralogical properties of sand and sandstone samples, and on the specific content of distinctive heavy minerals. Heavy minerals, in particular, acted as fingerprints for source areas and so could be used forensically to reconstruct the parent rocks in eroding source regions. Now, we use a battery of geochemical methods. But no matter how well we make this match between erosional source area and depositional sink, provenance studies cannot help us fully understand the dynamics of the sediment-routing system that conveyed it from source to sink. It is rather like being present at the birth of a baby and the funeral of the man, but missing out on the life story. To understand the life story requires insights into the functioning of sediment-routing systems geomorphically under tectonic and climatic forcing.

### Landscape-evolution models

A landscape-evolution model seeks to produce topography numerically in terms of the forcing mechanisms of climate and tectonics. Success is generally gauged by whether the resulting numerical landscape looks 'realistic'. The weakness is that the simulation of 'realistic' landscapes cannot be said to represent adequate model testing and validation, because the attainment of realism is conditional on the use of exponents and coefficients in the model equations for local erosion or deposition for which there may be weak independent support. These model equations are not like the governing equations of physics, but calibrated bulk parameterizations of observations. The issue of the extrapolation of local hydraulics or sediment dynamics to larger spatial scales and longer timescales is the classic problem of upscaling.

Let us take the example of the effects of cyclic glaciation, a mode of response to cyclic climate changes that Earth has experienced in the past few million years (see page 284). To build a numerical landscape-evolution model for times of glaciation we would need to know the sliding velocity of ice by solution of a chosen ice-dynamics equation, a proportionality constant in the ice-erosion equation that depends on the underlying rock type, a rheological law relating ice deformation to local stress, a model for ice accumulation and ablation, and knowledge of the temperature at the base of the ice. This might work theoretically by making a large number of assumptions, but the resulting model would be impossible to use in a simulation of Quaternary landscapes. Why? Because the necessary parameter values to inform a long-term landscape model are not currently available, and perhaps never will be. This humbling realization does not denigrate the efforts of modellers working at the human timescale, but instead prompts us to think afresh about what is required for success with upscaled models.

It is not immediately obvious that the factors controlling a long-term response may be different from those controlling local processes. To



**Figure 1 | The concept of a sediment-routing system.** Sediment is transferred from a source region to a sink along trajectories shown by the dashed lines. Some trajectories involve short transit times with brief periods of storage in the sediment-routing system (small circles) (for example, on the river-channel bed), whereas others involve long transit times with prolonged periods of storage (large circles) (for example, in bars and especially on flood plains). Storage of sediment implies buffering of incoming sediment flux signals.

illustrate this point, the factors controlling the rate of accumulation of sediment in a river hinge on the local gradient in sediment-transport rate, which is controlled by local hydraulic variables, the range of sediment available and the details of the local topography of the channel, bars, banks and flood plain. The factors controlling the long-term accumulation of sediment, on the other hand, are related to the prolonged realms of subsidence of Earth's surface that are controlled by geophysical parameters associated with the crust and mantle. The two sets of parameters, each correct in their own setting, could hardly be more different.

One way out of this fix is to reformulate local transport equations into new ones that can be directly constrained by observational data — data that are found in the geological record. The upscaled model therefore does not rely on knowledge of local hydraulic or sediment dynamics information that it is impossible to acquire from the geological recorder of past geomorphic-sedimentary processes<sup>7</sup>.

Currently, long-term numerical landscape evolution models lack predictive power because their rate parameters are poorly constrained, commonly being derived from restricted conditions at the human timescale, and making it difficult to justify their extrapolations in time and space. Response times are poorly known, varied and complex<sup>8</sup>, and more data on long-term response are clearly required.

### Measuring rates with dates

We need to make measurements of how Earth's surface has evolved over time spans long enough to capture the effects of both tectonics and climate, but what measurements should we make, and with what strategy? Earth scientists are faced with an imposing problem — to measure the erosional history of a landscape that no longer exists. But to do so is necessary if we are to understand the erosional engine that shapes Earth's surface. The study of the thermal history of rocks is the only method currently available to solve this problem.

Thermochronological techniques that capture the time-temperature trajectory of a rock<sup>9</sup>, such as apatite fission-track analysis and helium diffusion during U-Th radioactive decay, provide vital information on cooling attributable to the rise to the surface of the Earth of a rock or mineral during erosion. The cooling history is recorded over geological timescales, dependent on the critical temperature and methodology used. When thermochronological methods are used in combination, they have the potential to provide invaluable constraints on long-term erosional history, but with certain caveats. At shallow depths, a crystal's time-temperature history is likely to be influenced by temperature variations caused by the irregular topography and variable temperature of Earth's surface. And even the helium-dating technique has a temporal resolution that seems

clunky in relation to the fine scale of climate change. As a result, thermochronological methods, despite shedding much-needed light on long-term changes in the workings of the erosional engine, are unlikely to provide one-to-one connections between the high-frequency variations in climate that have typified the past few million years and landscape response. Other dating techniques, such as the use of nuclides produced during exposure of surfaces to cosmic radiation, offer a promising possibility of capturing this elusive landscape response to fine-scale climate change.

### Interactions and feedbacks

The topography caused by the formation of mountains perturbs atmospheric circulation and steers the jet stream, thereby directly influencing regional and local climatic patterns, such as the distribution of precipitation. Strong gradients in precipitation patterns in turn dictate erosional behaviour as well as ecosystem type. This first-order feedback between tectonics and climate, seen as major spatial variations in precipitation between the wet windward side and the dry lee, is uncontroversial. What is more contentious is that the impact of tiny raindrops, through erosion, might cause the localization of the mighty forces of tectonic deformation; that is, erosion over a tectonically deforming crust (Fig. 2) encourages a flux of rock towards the site of surface erosion. Consequently, it has been proposed that heavy monsoon rains, and the resulting high erosion rates, might cause dormant faults to become active, that earthquakes might be concentrated near areas of high erosion, and that growing kilometre-scale folds in the fronts of mountain belts might amplify rapidly once they have broken through Earth's surface and experience erosion. Surprisingly, rates of tectonic deformation near the surface of Earth, and earthquake risk, might therefore be influenced by climate.

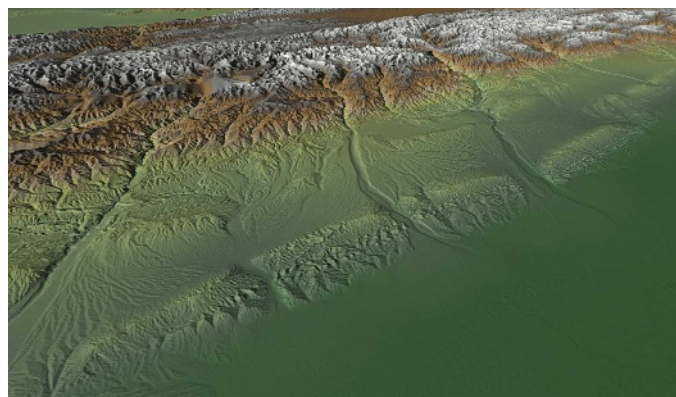
There is something of the chicken and egg in this debate, as there surely is in all strongly coupled systems. The same question was asked of the coupling between climate change and the surface uplift of mountain ranges in the past few million years<sup>10</sup>. Critical to answering the question of which triggers which is information on timing (so that connections can be made between different observational data sets) and information on system behaviour (so that the effects of internal dynamics can be understood and discriminated from external forcing such as climate change). As Jean Braun wrote<sup>11</sup>, "one must be reminded that the demonstration that a coupling between erosion and tectonics exists has so far been limited to the results of computer modelling of the system". So, what is known? At a basic level we know that high rates of deformation must be balanced by erosion, otherwise mountain ranges would continue to grow until they reached a height limit governed only by rock strength, or deep holes would appear in the core of mountain ranges in the absence of tectonic deformation. Observations from thermochronology, geochronology and sedimentation also make a close link in timing between erosional exhumation of mountain ranges and the fluxing of sediment onto fringing lowlands and ocean basins. Making sense of the more subtle couplings suggested by numerical models is now in the hands of observational geologists.

### The trap-door

Because of the intricate coupling between tectonic deformation and surface sediment-routing systems, tectonic deformation preserves the sediment in constant flux over Earth's surface<sup>12</sup>. Consequently, the realms of subsidence driven by tectonic processes<sup>13</sup> are the key to the transformation of erosional and depositional landscapes into the rock record of geological history.

Upland catchments release sediment into adjacent river systems or alluvial fans like a well-directed fire-hose, but the preservation of sediment in these systems depends on whether space is available for it to accumulate over long timescales. The amount of sediment preserved by this trap-door effect is generally minute compared with the overlying flux, but over geological timescales it is this small fraction that builds the sedimentary layers of stratigraphy. It is also the clue to why tectonics, rather than factors linked to surface sediment fluxes, controls long-term sediment accumulation rates.

In areas such as the Basin and Range province of the southwestern United States, where the crust is extending by slip along steep normal



**Figure 2 | Interaction of erosion, sedimentation and surface deformation.** Oblique-view digital elevation model of the Junggar region of central Asia, showing the interaction of erosion, sedimentation and deformation as the Tien Shan mountain range progrades tectonically into the sedimentary basin. The model is derived from 90-m resolution data from NASA's Shuttle Radar and Topography Mission. New tectonic folds are growing and being eroded in what was previously the low ground (green) of a sedimentary basin. (Image courtesy of K. Mueller, University of Colorado, Boulder.)

faults, it is believed that sediment fans, which accumulate against the uplifting mountain ranges, reflect the rate at which the faults slip by repeated earthquakes. Where the faults slip rapidly, the sediment fans are thought to have steep surface slopes and to show rapid down-system changes in particle size. The same effect can be seen in numerical models of much larger river systems. This fundamental control is invisible to the eyes of the observer standing on the surface. The critical parameter for a geomorphic trend is outside geomorphic space. There could hardly be a better advertisement for the integration of geomorphology and geology.

In essence, the burgeoning field of Earth surface processes requires a new conversation, so that the epic poem of Earth history can be better read and learned from. Figuratively, it requires atmospheric physicists to care about the tectonics of mountain ranges and for stratigraphers to care about fluvial hydrology. This new conversation will benefit from a close dovetailing of numerical modelling approaches with new observations relevant to a broad range of timescales.

Philip Allen is in the Department of Earth Science and Engineering, Imperial College, South Kensington Campus, London SW7 2AZ, UK.

1. Milliman, J. D. & Meade, R. H. Worldwide delivery of river sediment to the oceans. *J. Geol.* **91**, 1–21 (1983).
2. Squire, R. J., Campbell, I. H., Allen, C. M. & Wilson, C. J. L. Did the Transgondwanan Supermountain trigger the explosive radiation of animals on Earth? *Earth Planet. Sci. Lett.* **250**, 116–133 (2006).
3. Willett, S. D. Orogeny and orography: the effects of erosion on the structure of mountain belts. *J. Geophys. Res.* **104**, 28957–28981 (1999).
4. Allen, P. A. *Earth Surface Processes* (Blackwell, Oxford, 1997).
5. Métivier, Y. & Gaudemer, Y. Stability of output fluxes of large rivers in South and East Asia during the last 2 million years: implications for floodplain processes. *Basin Res.* **11**, 293–304 (1999).
6. Castellot, S. & Van Den Dreissche, J. How plausible are high-frequency sediment supply-driven cycles in the stratigraphic record? *Sediment. Geol.* **157**, 3–13 (2003).
7. Fedele, J. J. & Paola, C. Similarity solutions for fluvial sediment fining by selective deposition. *J. Geophys. Res. Earth Surf.* **112**, F02038, doi:10.1029/2005JF000409 (2007).
8. Allen, P. A. Striking a chord. *Nature* **434**, 961 (2005).
9. Braun, J., van der Beek, P. & Batt, G. *Quantitative Thermochronology: Numerical Methods for the Interpretation of Thermochronological Data* (Cambridge Univ. Press, Cambridge, 2006).
10. Molnar, P. & England, P. Late Cenozoic uplift of mountain ranges and global climate change: chicken and egg? *Nature* **346**, 29–34 (1990).
11. Braun, J. in *Analogue and Numerical Modelling of Crustal-Scale Processes* (eds Buiter, S. J. H. & Schreurs, G.) *Spec. Publ. Geol. Soc. Lond.* **253**, 307–325 (2006).
12. Leeder, M. R. Sedimentary basins: Tectonic recorders of sediment discharge from drainage catchments. *Earth Surf. Process. Landforms* **22**, 229–237 (1997).
13. Allen, P. A. & Allen, J. R. *Basin Analysis: Principles and Applications* 2nd edn (Blackwell, Oxford, 2005).
14. Malmton, D. V., Dunne, T. & Reneau, S. L. Stochastic theory of particle trajectories through alluvial valley floors. *J. Geol.* **111**, 525–542 (2003).

**Author Information** Reprints and permissions information is available at [npg.nature.com/reprints](http://npg.nature.com/reprints). Correspondence should be addressed to the author ([philip.allen@imperial.ac.uk](mailto:philip.allen@imperial.ac.uk)).

# The rise of atmospheric oxygen

Lee R. Kump

**Clues from ancient rocks are helping to produce a coherent picture of how Earth's atmosphere changed from one that was almost devoid of oxygen to one that is one-fifth oxygen.**

Imagine a *Star Trek* episode in which the Starship Enterprise stumbles into a time warp and is transported to Earth 3 billion years ago. The crew are eager to disembark but, before they do, they need to discover more about the pink methane haze<sup>1</sup> that surrounds the planet. The Starship Enterprise analyses a sample and, to the crew's surprise, it finds that Earth's atmosphere is as inhospitable as those of most of the celestial bodies they have encountered. Although the crew's hopes of exploring the surface of the early Earth are dashed, they did manage something that no one has done before. They determined the oxygen content of the early atmosphere.

## Timing is everything

Although it is probable that the history of atmospheric oxygen will be unravelled before the twenty-third century, which is when the television series *Star Trek* is set, more than 40 years of analysis of ancient rocks and of theoretical development have yet to produce a definitive picture of the planet's early history<sup>2</sup>. Two facts are known with certainty: Earth's earliest atmosphere was essentially devoid of oxygen; and today's atmosphere is composed of 21% oxygen. Most of the events that took place between these two time points are highly uncertain. By the end of the twentieth century, a battery of geological indicators suggested a shift from an anoxic to an oxic atmosphere some time between 2.5 and 2.0 billion years ago. This shift is known as the great oxidation event<sup>3</sup>. The most compelling evidence was the absence in older stratigraphic units of 'red beds', sedimentary rocks stained red by iron oxide. Instead, an abundance of lithified ancient soils that had lost their iron during weathering were found, reflecting the absence of oxygen in the weathering environment.

The 'smoking gun' for the rise of atmospheric oxygen was discovered and reported in 2000 (ref. 4). Rocks older than about 2.45 billion years contain a large degree of mass-independent fractionation (MIF) of sulphur isotopes; rocks younger than 2.32 billion years show essentially none<sup>5</sup> (Fig. 1). Many processes on Earth discriminate between the isotopes of elements, but usually the discrimination depends on the mass of the isotope. Processes that lead to MIF of sulphur are rare, and large MIF effects are restricted to gas-phase photochemical reactions in the upper atmosphere. The signature of MIF sulphur photochemistry is small and is rapidly homogenized in the modern oxidizing atmosphere. By contrast, in an oxygen-free atmosphere, large MIF effects are preserved, resulting in contrasting isotopic compositions of reduced and oxidized sulphur species that are deposited from the atmosphere and incorporated into sedimentary rocks.

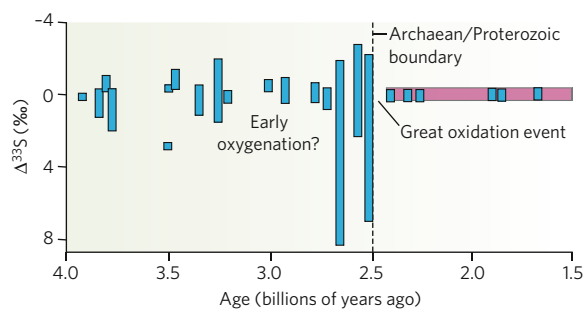
To preserve the MIF signature, three conditions are needed: very low atmospheric oxygen, sufficient sulphur gas in the atmosphere, and substantial concentrations of reducing gases. Numerical modelling by Zahnle *et al.*<sup>6</sup> has shown that the latter, in particular the atmospheric methane level, is the primary requirement for preserving MIF. Indeed, Zahnle *et al.* posit that the contraction of the spread in MIF values at ~2.45 billion years ago (Fig. 1) was the direct result of a collapse in atmospheric methane levels. This loss of 'greenhouse warming' is then invoked to explain the ensuing first major glaciation in Earth's history, perhaps of 'snowball Earth'

proportions<sup>7</sup> with ice extending to the tropics. In the scenario proposed by Zahnle *et al.*<sup>6</sup>, the decrease in methane would account for the increase in atmospheric oxygen, an alternative to the previously proposed scenario in which the rise in oxygen is proposed to have caused the collapse of the methane 'greenhouse'<sup>8</sup>. Given the high reactivity of methane and oxygen, the rise of oxygen and the demise of methane must have been inextricably linked; unravelling cause and effect will continue to be a challenge.

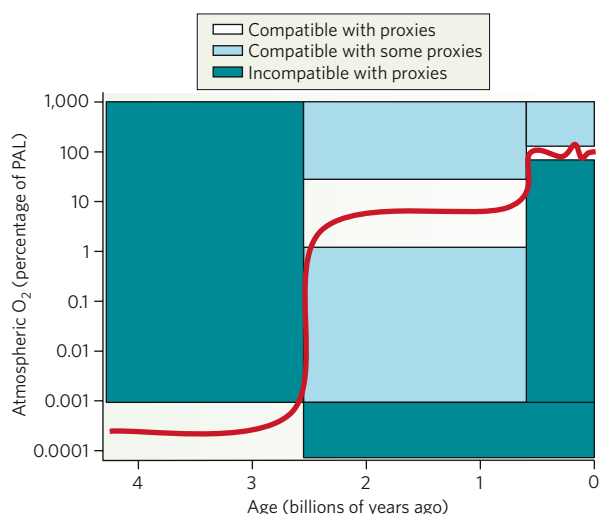
On closer inspection<sup>9,10</sup>, the Archaean (pre-2.5 billion years ago) MIF record displays an extended interval between 3.2 and 2.8 billion years ago (the Mesoarchaean) during which the spread of MIF values seems to be smaller. During this period<sup>11</sup>, was there a failed attempt at atmospheric oxygenation or a collapse of atmospheric methane, or is this simply an artefact of a sparse geological record? Trace gases, such as methane and carbon dioxide, are important in biogeochemical cycles, and their atmospheric concentrations have fluctuated significantly on geological timescales. Is it unreasonable to presume that when oxygen was a trace gas, it too varied substantially in response to imbalances between production and consumption? The most recent analysis of the MIF record indicates persistent anoxia throughout the Archaean, with some other change in atmospheric chemistry accounting for lower MIF values during the Mesoarchaean<sup>12</sup>, although geochemical evidence suggests a 'whiff of oxygen' might have appeared at the close of the Archaean, 50 million years before the permanent increase in oxygen<sup>13</sup>.

## From when to why

Future work on the evolution of atmospheric oxygen will focus on these intriguing aspects of the time before its ultimate rise at 2.45 billion years ago. It will seek to explain why the increase in oxygen occurred when it did, and to develop proxy indicators of oxygen levels so that the history of atmospheric oxygen evolution can be established.



**Figure 1 | Range of MIF of sulphur over time.** The great oxidation event occurred ~2.45 billion years ago, and an early, failed, oxygenation event might have occurred around 3.2 billion years ago (but this is hotly debated). The degree of MIF (blue) is indicated by  $\Delta^{33}\text{S}$ , which is the parts per thousand (‰) deviation of the standardized  $^{33}\text{S}/^{32}\text{S}$  ratio from the value predicted from the  $^{34}\text{S}/^{32}\text{S}$  ratio and mass-dependent fractionation. The range of values from samples of a given age is shown by vertical bars. The pink bar shows the range of variability in  $\Delta^{33}\text{S}$  that is due to mass-dependent effects, indicating only small variations during the past 2.32 billion years.



**Figure 2 | Prevailing view of atmospheric oxygen evolution over time.** The red line shows the inferred level of atmospheric oxygen bounded by the constraints imposed by the proxy record of atmospheric oxygen variation over Earth's history<sup>2,20</sup>. The signature of mass-independent sulphur-isotope behaviour sets an upper limit for oxygen levels before 2.45 billion years ago and a lower limit after that time. The record of oxidative weathering after 2.45 billion years ago sets a lower limit for oxygen levels at 1% of PAL, whereas an upper limit of 40% of PAL is inferred from the evidence for anoxic oceans during the Proterozoic. The tighter bounds on atmospheric oxygen from 420 million years ago to the present is set by the fairly continuous record of charcoal accumulation<sup>19</sup>: flames cannot be sustained below an oxygen level of 60% of PAL, and above about 160% of PAL the persistence of forest ecosystems would be unlikely because of the frequency and vigour of wildfires<sup>21</sup>.

Why oxygen levels rose when they did remains an understudied problem in atmospheric evolution. This time interval has traditionally been associated with the establishment of large, thick and stable continental land masses. Did a resultant change in the style of plate tectonics decrease the overall demand for oxygen as it reacted with volcanic<sup>14</sup> or metamorphic<sup>15</sup> outgasings? Or did cyanobacteria simply evolve oxygenic photosynthesis at this time<sup>7</sup>, perhaps in response to some new selective pressure arising from the stabilization of continents? Biomarker evidence for cyanobacteria (2-methylhopanes) and their waste product oxygen (in the form of steranes, which probably require oxygen for their synthesis) exists in rocks that formed 200 million years before the increase in atmospheric oxygen<sup>16</sup>. Taken together with other geological data, these biomarkers suggest that oxygen was being produced at prodigious rates before 2.5 billion years ago but was consumed faster than it was produced. However, 2-methylhopanes are no longer considered diagnostic of cyanobacteria<sup>17</sup>, and alternative pathways of sterane synthesis are possible<sup>1</sup>. So additional proxies must be sought. Fossilized microbial mats might hold the clue to the early origin of oxygen photosynthesis if it can be demonstrated that the expected strong redox gradients<sup>18</sup> existed and produced isotopic or compositional variations that can be recovered from Archaean rocks.

### Reconstructing ancient oxygen levels

Most geological indicators of ancient atmospheric oxygen levels imply only presence or absence (Fig. 2). MIF disappears when oxygen levels reach 0.001% of the present atmospheric level (PAL)<sup>8</sup>, and iron is retained in ancient lithified soils when oxygen is at 1% of its PAL<sup>3</sup>. Persistent anoxia of the oceans in the Proterozoic (from 1.8 to 0.5 billion

years ago) is argued to require oxygen levels below 40% of PAL<sup>2</sup>. Fire is sustained only above about 60% of PAL, so the more-or-less continuous geological record of charcoal over the past 450 million years sets this as a lower limit for atmospheric oxygen since the advent of forests on Earth. The interesting exception is the Middle to Late Devonian, ~380 million years ago, which shows a charcoal gap<sup>19</sup> coincident with widespread evidence for marine anoxia. The other available redox indicators are from marine sediments, requiring that internal ocean processes that affect deep-ocean oxygen levels be untangled before inferences about atmospheric oxygen level can be made.

A promising approach to reconstructing ancient oxygen levels looks at the effect that oxygen has on carbon-isotope fractionation<sup>20</sup>, but the signal is convolved with all the other factors that affect isotopic discrimination in plants and algae. Greater focus on the physiological effects of, adaptations to, and defences against oxygen in plants and animals is likely to lead to additional proxies. As we explore new proxies and seek out new sites for geological discovery, we will undoubtedly develop a more complete history of the multibillion-year evolution of atmospheric oxygen.

Lee R. Kump is in the Department of Geosciences, Pennsylvania State University, 535 Deike Building, University Park, Pennsylvania 16802, USA.

1. Lovelock, J. E. *The Ages of Gaia* (Norton, New York, 1988).
2. Canfield, D. E. The early history of atmospheric oxygen: homage to Robert M. Garrels. *Annu. Rev. Earth Planet. Sci.* **33**, 1–36 (2005).
3. Holland, H. D. in *Early Life on Earth* (ed. Bengtson, S.) 237–244 (Columbia Univ. Press, New York, 1994).
4. Farquhar, J., Bao, H. & Thiemens, M. Atmospheric influence of Earth's earliest sulfur cycle. *Science* **289**, 756–758 (2000).
5. Bekker, A., et al. Dating the rise of atmospheric oxygen. *Nature* **427**, 117–120 (2004).
6. Zahnle, K., Claire, M. & Catling, D. The loss of mass-independent fractionation in sulfur due to a Palaeoproterozoic collapse of atmospheric methane. *Geobiology* **4**, 271–283 (2006).
7. Kopp, R. E., Kirschvink, J. L., Hilburn, I. A. & Nash, C. Z. The paleoproterozoic snowball Earth: a climate disaster triggered by the evolution of oxygenic photosynthesis. *Proc. Natl Acad. Sci. USA* **102**, 11131–11136 (2005).
8. Pavlov, A. A. & Kasting, J. F. Mass-independent fractionation of sulfur isotopes in Archaean sediments: strong evidence for an anoxic Archaean atmosphere. *Astrobiology* **2**, 27–41 (2002).
9. Ono, S., Beukes, N. J., Rumble, D. & Fogel, M. L. Early evolution of atmospheric oxygen from multiple-sulfur and carbon isotope records of the 2.9 Ga Mozaan Group of the Pongola Supergroup, Southern Africa. *S. Afr. J. Geol.* **109**, 97–108 (2006).
10. Ohmoto, H., Watanabe, Y., Ikemi, H., Poulson, S. R. & Taylor, B. E. Sulphur isotope evidence for an oxic Archaean atmosphere. *Nature* **442**, 908–911 (2006).
11. Knauth, L. P. Signature required. *Nature* **442**, 873–874 (2006).
12. Farquhar, J. et al. Isotopic evidence for Mesoproterozoic anoxia and changing atmospheric sulfur chemistry. *Nature* **448**, 1033–1036 (2007).
13. Anbar, A. D. et al. A whiff of oxygen before the Great Oxidation Event? *Science* **317**, 1903–1906 (2007).
14. Kump, L. R. & Barley, M. E. Increased subaerial volcanism and the rise of atmospheric oxygen 2.5 billion years ago. *Nature* **448**, 1033–1036 (2007).
15. Catling, D. C. & Claire, M. W. How Earth's atmosphere evolved to an oxic state: a status report. *Earth Planet. Sci. Lett.* **237**, 1–20 (2005).
16. Brocks, J. J., Logan, G. A., Buick, R. & Summons, R. E. Archaean molecular fossils and the early rise of eukaryotes. *Science* **285**, 1033–1036 (1999).
17. Rashby, S. E., Sessions, A. L., Summons, R. E. & Newman, D. K. Biosynthesis of 2-methylbacteriohopanepolyols by an anoxygenic phototroph. *Proc. Natl Acad. Sci. USA* **104**, 15099–15104 (2007).
18. Herman, E. K. & Kump, L. R. Biogeochemistry of microbial mats under Precambrian environmental conditions: a modelling study. *Geobiology* **3**, 77–92 (2005).
19. Scott, A. C. & Glasspool, I. J. The diversification of Paleozoic fire systems and fluctuations in atmospheric oxygen concentration. *Proc. Natl Acad. Sci. USA* **103**, 10861–10865 (2006).
20. Berner, R. A., Beerling, D. J., Dudley, R., Robinson, J. M. & Wildman, R. A. Phanerozoic atmospheric oxygen. *Annu. Rev. Earth Planet. Sci.* **31**, 105–134 (2003).
21. Watson, A., Lovelock, J. E. & Margulis, L. Methanogenesis, fires, and the regulation of atmospheric oxygen. *Biosystems* **10**, 293–298 (1978).

**Acknowledgements** I thank the NASA Astrobiology Institute for supporting my research on atmospheric oxygen evolution.

**Author Information** Reprints and permissions are available at [npg.nature.com/reprints](http://npg.nature.com/reprints). Correspondence should be addressed to the author (lkump@psu.edu).

# An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics

James C. Zachos, Gerald R. Dickens & Richard E. Zeebe

**Past episodes of greenhouse warming provide insight into the coupling of climate and the carbon cycle and thus may help to predict the consequences of unabated carbon emissions in the future.**

By the year 2400, it is predicted that humans will have released about 5,000 gigatonnes of carbon (Gt C) to the atmosphere since the start of the industrial revolution if fossil-fuel emissions continue unabated and carbon-sequestration efforts remain at current levels<sup>1</sup>. This anthropogenic carbon input, predominantly carbon dioxide (CO<sub>2</sub>), would eventually return to the geosphere through the deposition of calcium carbonate and organic matter<sup>2</sup>. Over the coming millennium, however, most would accumulate in the atmosphere and ocean. Even if only 60% accumulated in the atmosphere, the partial pressure of CO<sub>2</sub> ( $p_{\text{CO}_2}$ ) would rise to 1,800 parts per million by volume (p.p.m.v.) (Fig. 1). A greater portion entering the ocean would decrease the atmospheric burden but with a consequence: significantly lower pH and carbonate ion concentrations of ocean surface layers<sup>1</sup> (Fig. 1).

A marked increase in atmospheric  $p_{\text{CO}_2}$  would increase mean global temperature, thereby affecting atmospheric and oceanic circulation, precipitation patterns and intensity, the coverage and thickness of sea ice, and continental ice-sheet stability. However, forecasting the timing and magnitude of these responses is challenging because they can be nonlinear. Of particular concern are potential positive feedbacks that could amplify increases in the concentrations of greenhouse gases — water, CO<sub>2</sub>, methane and nitrous oxide (N<sub>2</sub>O) — effectively escalating climate sensitivity to initial anthropogenic carbon input<sup>3</sup>. For example, ocean surface warming and freshwater discharge at high latitudes could slow the exchange of shallow and deep water in the ocean, impeding both abiotic and biotic removal of anthropogenic carbon from the atmosphere. Potential negative feedbacks are also garnering great interest. As a possible counterbalance to decreased density of surface water on a warmer Earth, stronger zonal winds might increase ocean overturning (see page 286).

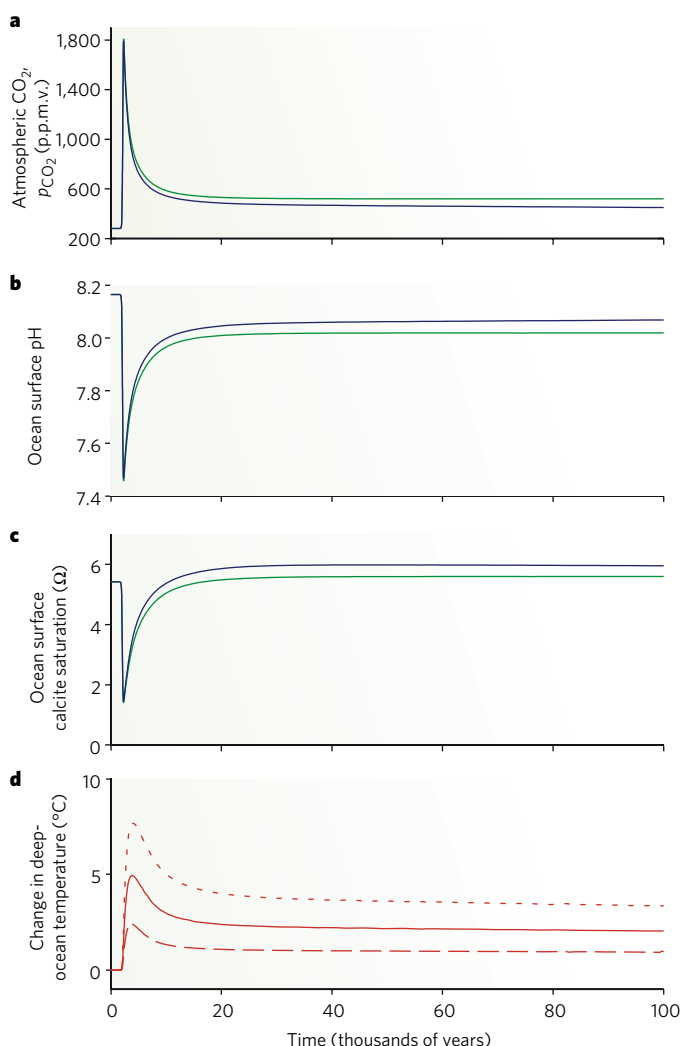
Observations of modern and Holocene (the past 10,000 years or so) climates have provided essential constraints for understanding climate dynamics and a baseline for predicting future responses to carbon input. But such observations can provide only limited insight into the response of climate to massive, rapid input of CO<sub>2</sub>. To evaluate climate theories more thoroughly, particularly with regard to feedbacks and climate sensitivity to  $p_{\text{CO}_2}$ , it is desirable to study samples obtained when CO<sub>2</sub> concentrations were high (approaching or exceeding 1,800 p.p.m.v.) and to make observations for intervals longer than those of ocean overturning and carbon cycling (more than 1,000 years)<sup>4</sup>. Earth scientists have therefore turned increasingly to ancient time intervals, particularly those in which  $p_{\text{CO}_2}$  was much higher than now, in which  $p_{\text{CO}_2}$  changed rapidly, or both. Recent reconstructions of Earth's history have considerably improved our knowledge of known 'greenhouse' periods and have uncovered several previously unknown episodes of rapid emissions of greenhouse gases and abrupt warming.

## Cenozoic greenhouse climates

The Cenozoic era, the last 65 million years of Earth's history, provides an ideal backdrop from which to understand relationships between carbon cycling and climate. In contrast to the present day, much of the early Cenozoic was characterized by noticeably higher concentrations of greenhouse gases, as well as a much warmer mean global temperature and poles with little or no ice<sup>5,6</sup> (Fig. 2). The extreme case is the Early Eocene Climatic Optimum (EECO), 51–53 million years ago, when  $p_{\text{CO}_2}$  was high and global temperature reached a long-term maximum. Only over the past 34 million years have CO<sub>2</sub> concentrations been low, temperatures relatively cool, and the poles glaciated. This long-term shift in Earth's climatic state resulted, in part, from differences in volcanic emissions, which were particularly high during parts of the Palaeocene and Eocene epochs (about 40–60 million years ago) but have diminished since then. Changes in chemical weathering of silicate rocks were also important<sup>7</sup>. On long timescales, this process sequesters CO<sub>2</sub>, preventing concentrations from rising too high or from falling too low. As the atmospheric CO<sub>2</sub> concentration rises, temperature and precipitation increase and thereby enhance chemical weathering; as the concentration declines, temperature and precipitation decrease, slowing weathering. Whereas other processes (such as the oxidation and burial of organic carbon) change CO<sub>2</sub> concentrations, the negative weathering feedback loop maintains Earth's climate within a habitable range over millions of years and longer<sup>7</sup>.

On shorter timescales, atmospheric CO<sub>2</sub> concentration and temperature can change rapidly, as demonstrated by a series of events during the early Cenozoic known as hyperthermals. These were relatively brief intervals (less than a few tens of thousands of years) of extreme global warmth and massive carbon addition but with widely differing scales of forcing and response. During the most prominent and best-studied hyperthermal, the Palaeocene–Eocene Thermal Maximum (PETM; about 55 million years ago), the global temperature increased by more than 5 °C in less than 10,000 years<sup>8</sup> (Fig. 3). At about the same time, more than 2,000 Gt C as CO<sub>2</sub> — comparable in magnitude to that which could occur over the coming centuries — entered the atmosphere and ocean.

Evidence for this carbon release is found in sedimentary records across the event. This includes a rapid and pronounced decrease in the <sup>13</sup>C/<sup>12</sup>C ratio of carbonate and organic carbon across the globe (that is, a negative carbon isotope excursion) and a prominent drop in the carbonate content of marine sediment deposited at several thousands of metres water depth (that is, a deep-sea dissolution horizon)<sup>8</sup>. The first observation indicates injection into the atmosphere or ocean of a very large mass of <sup>13</sup>C-depleted carbon, affecting the composition of the global carbon cycle. The second observation is a telltale signature of ocean acidification. The entire event lasted less than 170,000 years. Given the residence time of carbon (the average time a carbon atom spends in the ocean; about 100,000 years), this is consistent with a fast



**Figure 1 | Response to massive carbon input.** A simulation of atmospheric  $\text{CO}_2$  (a), ocean surface pH (b), ocean surface calcite saturation (c) and deep-ocean temperature changes (d) in response to the input of 5,000 Gt C of anthropogenic  $\text{CO}_2$  into the atmosphere, starting from pre-industrial  $\text{CO}_2$  levels (around the year 1860). These results were obtained with a carbon-cycle reservoir model coupled to a sediment model<sup>7,19</sup>. Blue and green curves indicate, respectively, runs with and without a silicate-weathering feedback. Silicate-weathering feedback involves the chemical dissolution of silicon-bearing rock on land, the primary permanent sink for  $\text{CO}_2$ . Projected changes in deep-ocean temperature in d assume a homogeneous warming of the ocean with a time lag of 1,000 years relative to atmospheric  $\text{CO}_2$  (ref. 2) and the following temperature sensitivities to a doubling of  $\text{CO}_2$  concentration: short-dashed line, 4.5  $^{\circ}\text{C}$ ; solid line, 3.0  $^{\circ}\text{C}$ ; long-dashed line, 1.5  $^{\circ}\text{C}$ .

release and subsequently slower removal of carbon. Several other early Eocene hyperthermals have been documented recently<sup>9</sup>, including the Eocene Thermal Maximum 2 (Fig. 2). Although their features have not yet been fully established, the events are also characterized by negative carbon isotope excursions and deep-sea carbonate dissolution horizons, but are proportionally smaller than for the PETM.

The source or sources of massive carbon injections during early Cenozoic hyperthermals remain uncertain. Carbon might have come from deeply buried rocks, perhaps liberated as methane and  $\text{CO}_2$  during intrusive volcanism<sup>10</sup>. Alternatively, it could have come from Earth's surface as a positive feedback to initial warming. For example, a rise in deep-sea temperature might have triggered the decomposition of gas hydrates on continental margins, releasing substantial amounts of methane and fuelling additional warming<sup>11</sup>. Another such source is the oxidation of organic matter in terrestrial environments<sup>12,13</sup>. In general, methane is appealing as a major source of carbon because it can be

markedly depleted in  $^{13}\text{C}$  and because it rapidly oxidizes to  $\text{CO}_2$  in the atmosphere and ocean.

Irrespective of source, the hyperthermals occurred over sufficiently short durations that plate tectonic boundary conditions, although different from those of the present day, did not change substantially. In this regard, the hyperthermals provide a special opportunity to investigate aspects of Earth-system dynamics operating 100–10,000 years after a massive injection of carbon. Already, recent studies of the PETM seem to validate some forecasts about future first-order changes in climate: extreme ocean warming of more than 5  $^{\circ}\text{C}$  extended to the North Pole; shifts in regional precipitation occurred, resulting in greater discharge from rivers at high latitudes and freshening of surface waters in the Arctic Ocean; and global ecosystems changed markedly, with major latitudinal and intercontinental migrations in terrestrial plants and mammals and with the sudden appearance of 'exotic' phytoplankton and zooplankton in open and coastal ocean environments (see refs 14 and 15 for reviews).

More importantly, the transient warming events show characteristics that are indicative of short-term positive feedbacks, which accelerated and magnified the effects of initial carbon injection before weathering and other negative feedbacks restored the global carbon cycle to a steady state. The most obvious characteristics are the timing and magnitude of various environmental signals. Stable-isotope and other records suggest that the abrupt and massive carbon input followed an interval of gradual warming and preceded an interval of decreased carbon uptake. Moreover, for the PETM, even the most conservative estimates of the mass of carbon released might require contributions from multiple sources.

### Opportunities and challenges

The overall conditions and transient hyperthermals of the early Cenozoic represent an assortment of natural experiments that can help researchers to investigate the coupling of carbon cycling and climate over a range of timescales, and thus provide a means of testing theory. Two important opportunities are to evaluate the role of physical and biogeochemical feedbacks in amplifying or moderating increases in concentrations of greenhouse gases, and to investigate the basic sensitivity of climate to extreme changes in concentrations of greenhouse gases.

### Feedbacks

The ocean is the primary carbon sink on moderate timescales (100–1,000 years), so of the 5,000 Gt C that humans could emit into the atmosphere (between the onset of the industrial revolution and the year 2400), the ocean would probably absorb roughly 70% after 1,000 years (Fig. 1). However, such carbon uptake depends on exchange between the thin and relatively warm surface layer that absorbs atmospheric  $\text{CO}_2$  and the much thicker and relatively cold deep-ocean reservoir that can store large amounts of carbon. As the small surface reservoir takes up  $\text{CO}_2$ , its pH decreases<sup>1</sup>, slowing the additional absorption of  $\text{CO}_2$ . To prevent the surface layer from becoming oversaturated, carbon must be shuttled quickly to the thermally isolated deep reservoir through advection (deep-ocean convection) or through the sinking of dead organisms (the biological pump). Unfortunately, rapid warming may compromise both processes. Warming and freshening of high-latitude surface water can slow the rate of convective overturning, and increased thermal stratification makes it more difficult for wind-driven mixing to return nutrients from the deep ocean to organisms in the photic zone (the upper 200 m or so of the water column, which is penetrated by sunlight, thereby allowing organisms to photosynthesize)<sup>3</sup>. Although such a state cannot be sustained indefinitely, because diffusive processes would transfer heat to the deep ocean, it could accelerate the increase in atmospheric  $\text{CO}_2$  concentrations relative to steady-state conditions.

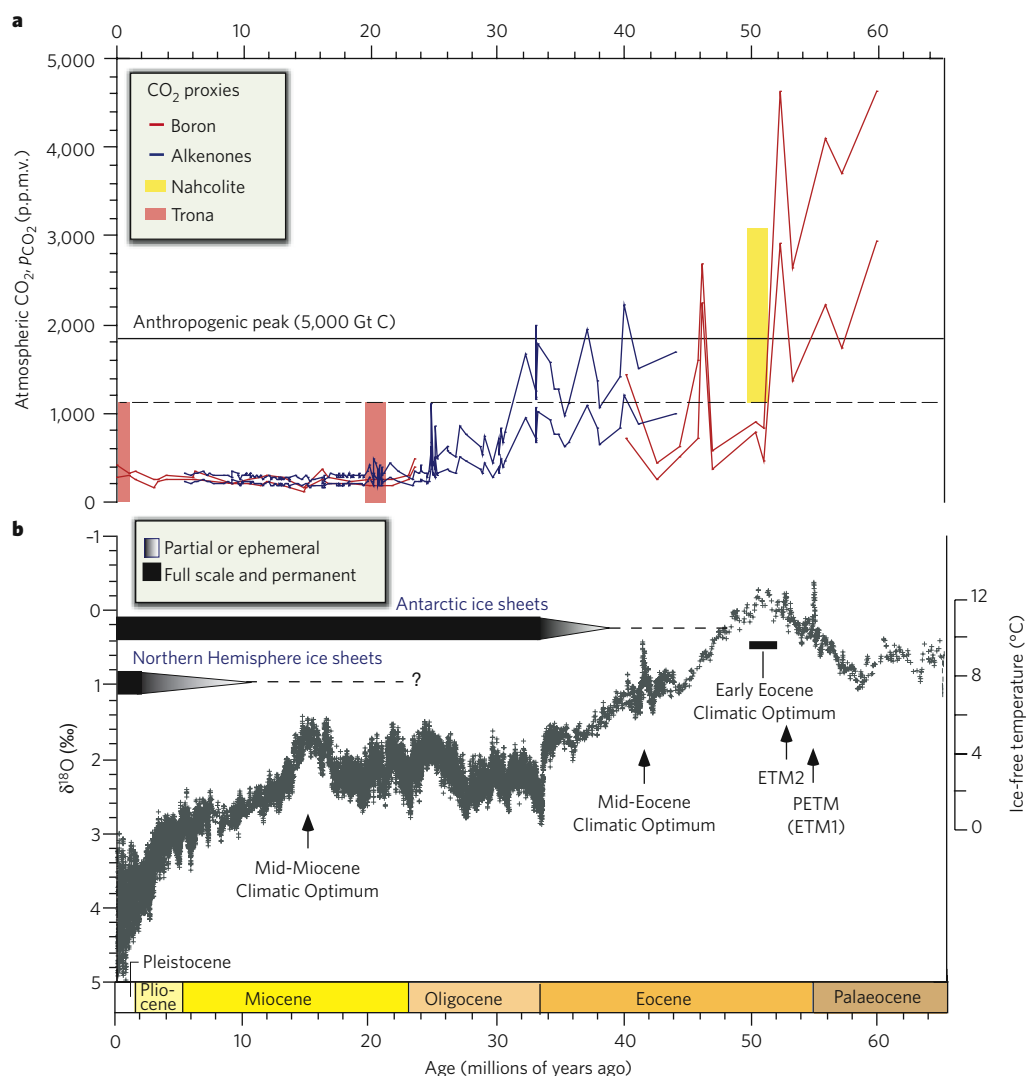
Is there any evidence of similar transient responses during past episodes of abrupt warming? High-resolution single-shell foraminiferal isotope records from the PETM suggest a delay of several thousand years in the propagation of the carbon isotope excursion from the surface ocean to the deep sea<sup>16</sup>, a pattern that could reflect a transient slowing of overturning circulation. If so, a combination of decreased

ocean overturning and increased surface temperatures should have decreased the flow of dissolved oxygen to deep water. Several direct lines of evidence, such as laminated sediment in cores from the Caribbean and central Arctic regions, suggest that dissolved oxygen did indeed decrease across the PETM. Moreover, the PETM coincided with a major extinction of benthic foraminiferans, with widespread oxygen deficiency in the ocean as a possible cause<sup>17</sup>.

With such ocean conditions, greater preservation and burial of solid organic carbon in deep-sea sediments might be predicted, effectively countering the decreased carbon flux from surface waters. However, this has not been documented. Two largely unexplored processes involving the microbial decomposition of organic carbon, both functioning as additional positive feedbacks, might operate during times of massive carbon input and rapid warming. Carbonate dissolution in the deep ocean decreases sedimentation rates, exposing organic carbon at or near

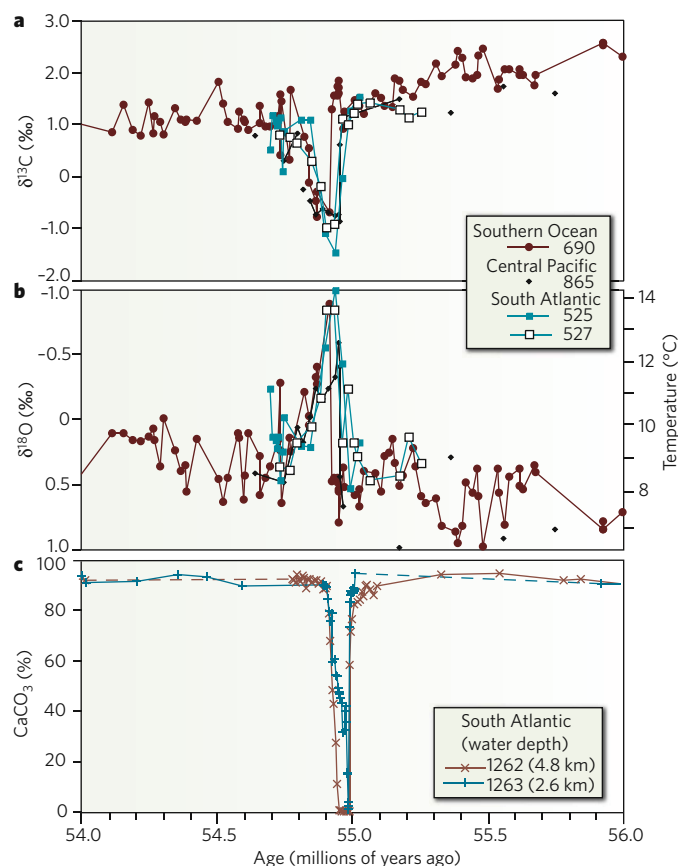
the sea floor for a longer duration, and warming of deep waters will accelerate overall microbial activity and the consumption of organic carbon. Future investigations might therefore focus specifically on the evidence for changes in ocean overturning, oxygen deficiency and the burial of organic carbon.

The positive feedbacks of greatest concern for understanding overall global warming may be those that could release hundreds to thousands of gigatonnes of carbon after initial warming<sup>11–13</sup>. The large masses of organic carbon stored in soils (for example, as peat) or sediments of shallow aquatic systems (for example, wetlands, bogs and swamps) represent a potential carbon input, should regions that were humid become drier. Rapid desiccation or fire could release carbon from these reservoirs at rates faster than carbon uptake by similar environments elsewhere. By contrast, regions that once were dry might emit methane as they become wetter<sup>18</sup>. Methane might also enter the ocean or atmosphere through the



**Figure 2 | Evolution of atmospheric CO<sub>2</sub> levels and global climate over the past 65 million years.** **a**, Cenozoic  $p_{\text{CO}_2}$  for the period 0 to 65 million years ago. Data are a compilation of marine (see ref. 5 for original sources) and lacustrine<sup>24</sup> proxy records. The dashed horizontal line represents the maximum  $p_{\text{CO}_2}$  for the Neogene (Miocene to present) and the minimum  $p_{\text{CO}_2}$  for the early Eocene (1,125 p.p.m.v.), as constrained by calculations of equilibrium with Na–CO<sub>3</sub> mineral phases (vertical bars, where the length of the bars indicates the range of  $p_{\text{CO}_2}$  over which the mineral phases are stable) that are found in Neogene and early Eocene lacustrine deposits<sup>24</sup>. The vertical distance between the upper and lower coloured lines shows the range of uncertainty for the alkenone and boron proxies. **b**, The climate for the same period (0 to 65 million years ago). The climate curve is a stacked deep-sea benthic foraminiferal oxygen-isotope curve based on records from

Deep Sea Drilling Project and Ocean Drilling Program sites<sup>6</sup>, updated with high-resolution records for the interval spanning the middle Eocene to the middle Miocene<sup>25–27</sup>. Because the temporal and spatial distribution of records used in the stack are uneven, resulting in some biasing, the raw data were smoothed by using a five-point running mean. The  $\delta^{18}\text{O}$  temperature scale, on the right axis, was computed on the assumption of an ice-free ocean; it therefore applies only to the time preceding the onset of large-scale glaciation on Antarctica (about 35 million years ago). The figure clearly shows the 2-million-year-long Early Eocene Climatic Optimum and the more transient Mid-Eocene Climatic Optimum, and the very short-lived early Eocene hyperthermals such as the PETM (also known as Eocene Thermal Maximum 1, ETM1) and Eocene Thermal Maximum 2 (ETM2; also known as ELMO). %, parts per thousand.



**Figure 3 | Low-resolution marine stable-isotope records of the PETM and the carbon isotope excursion, together with the seafloor sediment  $\text{CaCO}_3$  record.** The carbon isotope (a) and oxygen isotope (b) records are based on benthic foraminiferal records (see ref. 6 for original sources), and the  $\text{CaCO}_3$  records (c) are from drill holes in the South Atlantic<sup>8</sup>. Panel b also shows inferred temperatures. Ocean drilling site locations are indicated in the keys. The decrease in sedimentary  $\text{CaCO}_3$  reflects increased dissolution and indicates a severe decrease in seawater pH (that is, ocean acidification). The base of the  $\text{CaCO}_3$  dissolution horizon is below the onset of the carbon isotope excursion because most of the carbonate dissolution involved uppermost Palaeocene sediments that were deposited before the event (chemical erosion). Panels a and b adapted, with permission, from ref. 6.

dissociation of gas hydrate in marine sediment. This feedback would probably take several thousands of years to initiate because heat must be propagated by ocean advection to water depths at which hydrates can form (more than 1 km in the early Cenozoic), and then by diffusion into sediments. However, the amount of methane that could be liberated is enormous, and after gas hydrate dissociation was initiated, the flux might proceed rapidly as overpressured pore waters triggered fluid expulsion or sediment slides on the sea floor<sup>11</sup>.

These potential carbon-cycle feedbacks for amplifying warmth are not fully understood. In fact, demonstrating that such feedbacks have operated in the past remains a major challenge. The abrupt negative carbon isotope excursions that mark the hyperthermals and attest to a massive input of isotopically depleted carbon cannot be used alone to identify the source, especially if more than one source existed. Records of geochemical or physical fingerprints, such as hopanoids from methanotrophs or charcoal from wildfires, would help<sup>18</sup>. Constraining the rate and mass of carbon released, for example by quantifying changes in ocean carbonate chemistry, is also essential for identifying sources<sup>19</sup>.

The PETM and other hyperthermals should also provide insight into the longer-term response of the carbon cycle to massive inputs of carbon, including the primary negative feedbacks that temporarily and permanently sequester carbon. Various simulations of the long-term fate of anthropogenic carbon emissions show consistent results. After

the cessation of emissions, and a peak in atmospheric  $p_{\text{CO}_2}$ , the ocean steadily absorbs much of the carbon, although with a decrease in pH and carbonate-ion concentration (Fig. 1). The carbonate-ion concentration is restored by dissolution of carbonate on the sea floor within several thousand years, but dissolved inorganic carbon and alkalinity remain high for tens of thousands of years afterwards. As a consequence, atmospheric  $p_{\text{CO}_2}$  does not return to pre-anthropogenic values but stabilizes at levels at least 50% higher than before the carbon injection (Fig. 1).

Marine-sediment records that span the PETM show features consistent with this pattern. The initial release of carbon, as represented by the carbon isotope excursion, is accompanied by widespread and significant dissolution of seafloor carbonate and a net deficit in deep-sea carbonate accumulation (Fig. 3). This is followed by an increase in carbonate accumulation at many locations, presumably reflecting a recovery of carbonate ion concentration<sup>8</sup>. Interestingly, carbonate accumulation during this recovery phase seems greater than before carbon injection, suggesting carbonate oversaturation. Although no detailed reconstructions of  $p_{\text{CO}_2}$  are available for the PETM, surface temperatures remain warm for thousands of years after the input of carbon seems to cease. Thus, at first glance, observations of the PETM support the theory about the long-term fate of fossil-fuel  $\text{CO}_2$ . The carbonate 'overshoot' represents a negative feedback, probably through enhanced silicate weathering and delivery of dissolved calcium and bicarbonate to the ocean. Gauging the sensitivity of this effect will enable the establishment of constraints on long-term forecasts for the carbon cycle following anthropogenic carbon emissions.

### Climate sensitivity

Early Cenozoic climate has received considerable interest because the response of climate to a broad range of high atmospheric values of  $p_{\text{CO}_2}$  (probably 1,000 to more than 2,000 p.p.m.v.) can be examined. One feature common to all greenhouse periods, whether transient or long-lived, is exceptionally warm poles<sup>15</sup>. In the more extreme cases, the EECO and PETM, high-latitude temperatures were substantially higher than can be simulated by models without unreasonably high  $p_{\text{CO}_2}$  (refs 20, 21). Somehow, models are not precisely simulating processes critical to poleward heat transport, albedo, or polar heat retention at higher greenhouse gas levels. Modified ocean heat transport has been investigated and found to be incapable of transporting heat fast enough to compensate for polar heat loss<sup>22</sup>. In contrast, polar stratospheric clouds, which might have been more extensive during the greenhouse intervals because of higher concentrations of methane in the atmosphere, seem to be very effective at trapping heat<sup>20</sup>. Similarly, non- $\text{CO}_2$  greenhouse gases, which are usually neglected, may have had a major role. Recent theoretical and experimental studies indicate that, under high  $p_{\text{CO}_2}$ , background concentrations of trace gases such as methane and  $\text{N}_2\text{O}$  should be higher because of greater production under warmer and wetter conditions (that is, more extensive wetlands) and because of lower rates of oxidation in the atmosphere (resulting from lower emissions of volatile organic compounds by plants)<sup>21</sup>. Collectively, such physical and biochemical feedbacks would tend to enhance the sensitivity of climate to changes in  $\text{CO}_2$  and might explain the unusual polar warmth of the early Cenozoic.

Another prominent feature of the transient greenhouse episodes, specifically the PETM, are marked shifts in the distribution and intensity of precipitation, as inferred from fossil vegetation and other proxy data. Most regions, particularly in middle to high latitudes, experienced a shift towards wetter climates. However, the response on a regional scale was far more complex. For example, recent studies show that some regions, such as the western interior of North America, became drier at the onset of the PETM, whereas other regions, such as western Europe, experienced increased extreme precipitation events and massive flooding<sup>23</sup>. These palaeo-observations imply a high degree of sensitivity in the hydrological cycle to extreme changes in  $p_{\text{CO}_2}$  and temperature. Additional documentation of precipitation changes for climatically sensitive regions during Eocene greenhouse episodes could prove useful for assessing how well models simulate extremes in climate.

## Outlook for the future

If fossil-fuel emissions continue unabated, in less than 300 years  $p_{\text{CO}_2}$  will reach about 1,800 p.p.m.v., a level not present on Earth for roughly 50 million years. Both the magnitude and the rate of rise complicate the goal of accurately forecasting how the climate will respond. Foremost among the challenges that must be overcome to achieve this goal is the development of a deeper understanding of the complex interactions that link the climate system with the biogeochemical cycles, specifically the role of positive and negative feedbacks. The occurrence of past greenhouse warming events provides one opportunity to test theory about the physical and biogeochemical interactions in rapidly shifting systems. There are of course limitations on which facets of theory and models can be tested given uncertainties in proxies and the limited spatial and temporal resolution of palaeorecords. Nevertheless, the past greenhouse events provide glimpses of the future. Until the most salient features of these events, for example the global patterns of carbonate deposition or the extreme polar warmth, can be replicated with dynamical models, forecasts of climate beyond the next century (that is, under extreme greenhouse gas levels) should be viewed with caution, and efforts to comprehend the underlying physics and biogeochemistry of the coupling between climate and the carbon cycle should be hastened. ■

James C. Zachos is in the Department of Earth and Planetary Sciences, University of California at Santa Cruz, Santa Cruz, California 95060, USA. Gerald R. Dickens is in the Department of Earth Sciences, Rice University, Houston, Texas 77005, USA. Richard E. Zeebe is at the School of Ocean and Earth Science and Technology, University of Hawaii at Manoa, 1000 Pope Road, MSB 504, Honolulu, Hawaii 96822, USA.

- Caldeira, K. & Wicket, M. E. Anthropogenic carbon and ocean pH. *Nature* **425**, 365–365 (2003).
- Archer, D. Fate of fossil fuel  $\text{CO}_2$  in geologic time. *J. Geophys. Res. Oceans* **110**, C09S05, doi:10.1029/2004JC002625 (2005).
- Friedlingstein, P. et al. Climate–carbon cycle feedback analysis: Results from the (CMIP)-M-4 model intercomparison. *J. Clim.* **19**, 3337–3353 (2006).
- Doney, S. C. & Schimel, D. S. Carbon and climate system coupling on timescales from the Precambrian to the Anthropocene. *Annu. Rev. Environ. Resources* **32**, 14.1–14.36 (2007).
- Royer, D. L.  $\text{CO}_2$ -forced climate thresholds during the Phanerozoic. *Geochim. Cosmochim. Acta* **70**, 5665–5675 (2006).
- Zachos, J., Pagani, M., Sloan, L., Thomas, E. & Billups, K. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**, 686–693 (2001).
- Walker, J. C. G., Hays, P. B. & Kasting, J. F. A negative feedback mechanism for the long-term stabilization of Earth's surface-temperature. *J. Geophys. Res. Oceans Atmos.* **86**, 9776–9782 (1981).
- Zachos, J. C. et al. Rapid acidification of the ocean during the Paleocene–Eocene Thermal Maximum. *Science* **308**, 1611–1615 (2005).
- Lourens, L. J. et al. Astronomical pacing of late Palaeocene to early Eocene global warming events. *Nature* **435**, 1083–1087 (2005).
- Svensen, H. et al. Release of methane from a volcanic basin as a mechanism for initial Eocene global warming. *Nature* **429**, 524–527 (2004).
- Dickens, G. R. Rethinking the global carbon cycle with a large, dynamic and microbially mediated gas hydrate capacitor. *Earth Planet. Sci. Lett.* **213**, 169–183 (2003).
- Kurtz, A. C., Kump, L. R., Arthur, M. A., Zachos, J. C. & Paytan, A. Early Cenozoic decoupling of the global carbon and sulfur cycles. *Paleoceanography* **18**, 1090, doi:10.1029/2003PA000908 (2003).
- Higgins, J. A. & Schrag, D. P. Beyond methane: Towards a theory for the Paleocene–Eocene Thermal Maximum. *Earth Planet. Sci. Lett.* **245**, 523–537 (2006).
- Wing, S. L., Gingerich, P. D., Schmitz, B. & Thomas, E. (eds). *Causes and Consequences of Globally Warm Climates in the Early Paleocene* (Geol. Soc. Am. Spec. Pap. 369, Boulder, Colorado, 2003).
- Sluijs, A., Bowen, G. J., Brinkhuis, H., Lourens, L. J. & Thomas, E. in *Deep-Time Perspectives on Climate Change: Marrying the Signal from Computer Models and Biological Proxies* (eds Williams, M. et al.) 323–349 (Geological Society of London, London, 2007).
- Thomas, D. J., Zachos, J. C., Bralower, T. J., Thomas, E. & Bohaty, S. Warming the fuel for the fire: Evidence for the thermal dissociation of methane hydrate during the Paleocene–Eocene Thermal Maximum. *Geology* **30**, 1067–1070 (2002).
- Thomas, E. & Shackleton, N. J. in *Correlation of the Early Paleogene in Northwest Europe* (eds Knox, R. W. O. B., Corfield, R. M. & Dunay, R. E.) 401–441 (Geol. Soc. Lond. Spec. Publ. 101, London, 1996).
- Pancost, R. D. et al. Increased terrestrial methane cycling at the Palaeocene–Eocene Thermal Maximum. *Nature* **449**, 332–335 (2007).
- Zeebe, R. E. & Zachos, J. C. Reversed deep-sea carbonate ion basin gradient during Paleocene–Eocene Thermal Maximum. *Paleoceanography* **22**, PA3201, doi:10.1029/2006PA001395 (2007).
- Sloan, L. C. & Pollard, D. Polar stratospheric clouds: A high latitude warming mechanism in an ancient greenhouse world. *Geophys. Res. Lett.* **25**, 3517–3520 (1998).
- Beerling, D. J., Hewitt, C. N., Pyle, J. A. & Raven, J. A. Critical issues in trace gas biogeochemistry and global change. *Phil. Trans. R. Soc. A* **365**, 1629–1642 (2007).
- Huber, M. & Sloan, L. C. Heat transport, deep waters, and thermal gradients: Coupled simulation of an Eocene greenhouse climate. *Geophys. Res. Lett.* **28**, 3481–3484 (2001).
- Schmitz, B. & Pujalte, V. Abrupt increase in seasonal extreme precipitation at the Paleocene–Eocene boundary. *Geology* **35**, 215–218 (2007).
- Lowenstein, T. K. & Demicco, R. V. Elevated Eocene atmospheric  $\text{CO}_2$  and its subsequent decline. *Science* **313**, 1928–1928 (2006).
- Billups, K., Channell, J. E. T. & Zachos, J. Late Oligocene to early Miocene geochronology and paleoceanography from the subantarctic South Atlantic. *Paleoceanography* **17**, U39–U49 (2002).
- Bohaty, S. M. & Zachos, J. C. Significant Southern Ocean warming event in the late middle Eocene. *Geology* **31**, 1017–1020 (2003).
- Palike, H. et al. The heartbeat of the Oligocene climate system. *Science* **314**, 1894–1898 (2006).

**Author Information** Reprints and permissions information is available at [npj.nature.com/reprints](http://npj.nature.com/reprints). Correspondence should be addressed to J.C.Z. and R.E.Z. ([jzachos@es.ucsc.edu](mailto:jzachos@es.ucsc.edu); [zeebe@hawaii.edu](mailto:zeebe@hawaii.edu)).

# Unlocking the mysteries of the ice ages

Maureen E. Raymo & Peter Huybers

**Much progress has been made towards understanding what caused the waxing and the waning of the great ice sheets, but a complete theory of the ice ages is still elusive.**

Perhaps the longest-standing puzzle in the Earth sciences is what caused the Northern Hemisphere ice sheets to come and go. Earth scientists have been trying to solve this puzzle since 1840, when Louis Agassiz proposed that the geological deposits in Europe and North America were the remnants of vast ice sheets that spilled from the mountains.

Joseph Adhémar seems to have been the first to suggest that glaciation was associated with changes in the configuration of Earth's orbit relative to the Sun. In 1842, he proposed that glaciation occurs when winters are anomalously long, which happens when they coincide with aphelion (the point of Earth's orbit that is farthest from the Sun). James Croll subsequently argued in the 1860s that glaciation occurs when winters coincide with aphelion not because such winters are longer but because the intensity of insolation (that is, solar radiation) is weaker at this point. At present, the favoured hypothesis is that proposed by Milutin Milanković, who turned Croll's argument on its head in the 1930s. He argued that glaciation occurs when insolation intensity is weak at high northern latitudes during summer. This happens when both Earth's spin axis is less tilted with respect to the orbital plane and aphelion coincides with summer (not winter) in the Northern Hemisphere. According to Milanković, when there is less insolation during the summer, snow and ice persist through the year, gradually accumulating into an ice sheet.

In 1976, James Hays, John Imbrie and Nicholas Shackleton<sup>1</sup> unearthed strong evidence in support of the orbital hypothesis of glaciation. Applying the newly developed geomagnetic timescale to a deep-sea sediment core, they showed that long-term variations in oxygen isotope ratios, as recorded in fossils of foraminifera, were concentrated at the frequencies predicted by the orbital hypothesis. The ratio of oxygen-18 to oxygen-16 ( $\delta^{18}\text{O}$ ) in the ocean was known to increase with glaciation, because oxygen-16 evaporates preferentially and is concentrated in ice sheets. Hays *et al.*<sup>1</sup> showed that  $\delta^{18}\text{O}$  varied with cycles of 41,000 years, the period associated with changes in the tilt of Earth's spin axis (or obliquity), and around 21,000 years, the period associated with the location of aphelion with respect to the seasons (also known as climatic precession or the precession of the equinoxes). But other questions immediately arose. The authors also found that, during the past 800,000 years, ice sheets took about 90,000 years to grow and only 10,000 years to collapse. They proposed a link with the eccentricity of Earth's orbit, which varies at periods of about 100,000 years. Earth's eccentricity has only a weak effect on incoming solar radiation, however, so the strong presence of a 100,000-year cycle was perplexing. Likewise, it was unclear why the rates of growth and collapse were asymmetrical.

Since this study was published, and with improvements in the dating of geological samples, strong evidence for an orbital influence on climate has been found across the globe. An understanding of how Earth's insolation has varied in the past and observations of the subsequent shifts in climate provide an opportunity to probe the mechanisms that control long-term climate change. Several hurdles must be overcome, however, before this knowledge can be used to its full potential. The climate physics and chemistry that are best understood are mainly attuned to processes that occur at daily to interannual timescales. Are the important factors

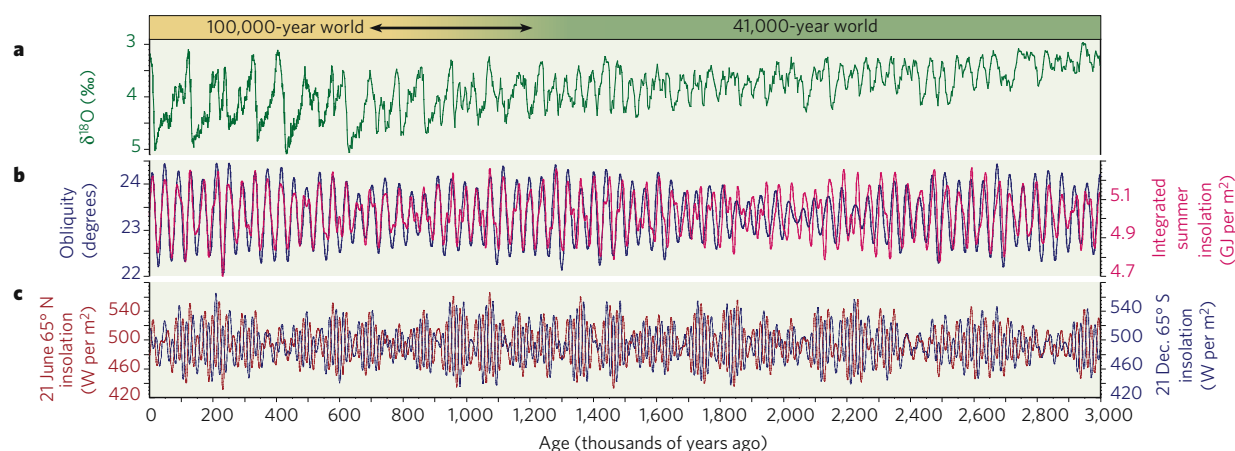
that regulate climate over centennial and longer timescales known? When a climate model is stepped forward, by minutes or days at a time, for hundreds or thousands of years, are the final results realistic? Climate scientists still do not understand how the subtle shifts in insolation at the top of the atmosphere are converted into massive changes in the ice volume on the ground.

## Regular timing

To tackle these problems, some researchers have turned to a time when glaciation seems to have been relatively straightforward. The glacial cycles of the late Pliocene to early Pleistocene (~1–3 million years ago) were more regular than those of the late Pleistocene, typically lasting about 41,000 years (Fig. 1a), which matches the period of change in Earth's tilt<sup>2</sup> (Fig. 1b). But how is the lack of variability with respect to precession explained? Precession, which occurs mainly at 23,000-year and 19,000-year intervals, is the orbital component that most influences summer insolation intensity (Fig. 1c). Indeed, precession is clearly observed in the ice-volume and sea-level records for the past 700,000 years. The few computer models that have been used to study the climate history of the late Pliocene to early Pleistocene also show a strong precession signal in the modelled ice volume. Are these climate models missing a fundamental piece of climate or ice-sheet physics, or are the assumptions about ice-volume proxies, such as  $\delta^{18}\text{O}$ , flawed? Both are possibilities.

One resolution to the puzzle of the missing precession variance harks back to Adhémar's proposal: summers with the greatest insolation intensity are also about a week shorter than the average duration of summer, because Earth orbits more quickly when it is close to the Sun. Peter Huybers<sup>3</sup> proposed that the amount of melting an ice sheet undergoes is better gauged by integrating the total insolation over summer (with summer defined as the period when insolation intensity exceeds a melting threshold), as opposed to using either the peak or the mean intensity of summer insolation. This summer-energy metric varies mainly at the obliquity period and is therefore consistent with the oxygen isotope record observed in marine samples from 1–3 million years ago (Fig. 1a, b). Why, then, do ice-sheet computer models, which implicitly incorporate integrated insolation and a full seasonal cycle, show that precession is the dominant control on the amount and timing of ablation? Huybers and Eli Tziperman<sup>4</sup> recently demonstrated that an ice-sheet model can generate 40,000-year glacial cycles when two conditions are met: the zone where the ice sheet is ablated must be north of about 60° N, where changes in obliquity have a greater effect on insolation, and the summer melt season must be long enough for changes in its duration to balance changes in insolation intensity. In the model of Huybers and Tziperman, a longer melt season results from sliding at the base of the ice sheet, which thins the ice sheet and draws its surface down to lower (warmer) elevations. They proposed that these factors are responsible for the different behaviour of the earlier ice sheets.

Climate modellers, geologists and geochemists should take note. These are testable predictions that are reminiscent of the regolith hypothesis of Peter Clark and David Pollard<sup>5</sup>. To reconcile the observation of late



**Figure 1 | Ice-age climate and solar variability.** A 3-million-year record of  $\delta^{18}\text{O}$  (ref. 8) (a); orbital obliquity (blue) compared with integrated summer insolation (red)<sup>3</sup> (b); and summer insolation for the Northern Hemisphere (on 21 June at 65° N; red) and the Southern Hemisphere (on 21 December at 65° S; blue)<sup>6</sup> (c).  $\delta^{18}\text{O}$  is considered a proxy of global ice-volume change, which is assumed to occur mostly in the Northern Hemisphere over this

interval. From 3 to 1 million years ago,  $\delta^{18}\text{O}$  varies primarily at the 41,000-year period characteristic of obliquity and integrated insolation. From 1 million years ago to the present, longer cycles of climate change, with a roughly 100,000-year period, are more obvious. The double-headed arrow indicates a transition more gradual than abrupt over the time indicated. GJ, gigajoule; W, watts.

Pliocene to early Pleistocene ice-margin deposits in Iowa and Kansas (at 40° N) during a time when the marine oxygen isotope record (Fig. 1a) suggests that ice sheets were smaller, Clark and Pollard proposed that a glacial substrate of easily deformable sedimentary rocks allowed basal sliding to increase and therefore resulted in a continental ice sheet that was thinner overall. They proposed that the gradual erosion of this upper sedimentary layer by ice sheets led to the transition to the larger, less mobile ice sheets of the late Pleistocene that varied at the slower 100,000-year periodicity. Was the maximum extent of the ice edge typically as far south as 40° N between 1 and 3 million years ago? Can sedimentological and mineralogical evidence be found for a long-term change in the erosional substrate scoured by this ice sheet? How thick were the late Pliocene to early Pleistocene ice sheets on North America? Are the proposed changes in basal sliding realistic? The answers to these questions have important implications for climate models.

### Out of phase

Another explanation for the lack of a precession signal in records of ice volume was proposed by Maureen Raymo, Lorraine Lisiecki and Kerim Nisancioglu<sup>6</sup>. They put forward a model in which Northern Hemisphere ice sheets wax and wane at precession periods, driven by the strongly nonlinear response of ice ablation to summer insolation intensity. In this model, however, the precession component of changes in ice volume is missing from marine records of  $\delta^{18}\text{O}$  because it is 'cancelled out' by changes in Southern Hemisphere ice volume that are of opposite phase. The effect of the precession of the equinoxes on summer insolation intensity is out of phase between hemispheres, whereas the effect of obliquity is in phase (Fig. 1c; look at times when precession is weak, such as ~2.4 million years ago). Thus, precession-paced changes in ice volume in each hemisphere would cancel out in globally integrated proxies such as ocean  $\delta^{18}\text{O}$  or sea level, leaving the in-phase obliquity (41,000-year) component of ice volume to dominate the records. Even a few tens of metres of ice-volume variance in the Southern Hemisphere would be enough to effectively hide a much greater Northern Hemisphere precession signal.

Is this possible? Could a terrestrial ice margin sensitive to local summer insolation have waxed and waned on East and West Antarctica at that time in the late Pliocene and early Pleistocene? We know little about the history of Antarctica at that time. The Antarctic Geological Drilling (ANDRILL) programme has astonished scientists recently with evidence for periodic warm open waters in the Ross Sea up until as recently as 1 million years ago<sup>7</sup>. And any evidence for a terrestrial ice margin at that time is now buried under the marine-based margin that encircles East Antarctica. To test

this idea for the origin of the '41,000-year world', well-dated proxy records sensitive to local climate and to the lateral movement of ice margins on land (in both the Northern Hemisphere and the Southern Hemisphere) are needed. Will such records show precession pacing? Similarly, Antarctic ice cores that extend into the early Pleistocene would help to determine whether, at that time, the local climate was in phase (as it is today) or out of phase with Northern Hemisphere insolation changes. Planning for such expeditions is already under way in the ice-core community. It could be that the East and West Antarctic ice sheets have had a far more dynamic history than has been thought.

It is widely accepted that variations in Earth's orbit affect glaciation, but a better and more detailed understanding of this process is needed. How can the 41,000-year glacial cycles of the early Pleistocene be explained, let alone the ~100,000-year glacial cycles of the late Pleistocene? How do the subtle changes in insolation relate to the massive changes in climate known as glacial cycles? And what are proxy climate records actually measuring? The field now faces these important questions, which are made all the more pressing as the fate of Earth's climate is inexorably tied to the vestige of Northern Hemisphere glaciation that sits atop Greenland, and to its uncertain counterpart to the south.

Maureen Raymo is in the Department of Earth Sciences, Boston University, 685 Commonwealth Avenue, Boston, Massachusetts 02215, USA. Peter Huybers is in the Department of Earth and Planetary Sciences, Harvard University, 20 Oxford Street, Cambridge, Massachusetts 02138, USA.

1. Hays, J. D., Imbrie, J. & Shackleton, N. J. Variations in the Earth's orbit: pacemaker of the ice ages. *Science* **194**, 1121–1131 (1976).
2. Pisias, N. G. & Moore, T. C. The evolution of Pleistocene climate: a time series approach. *Earth Planet. Sci. Lett.* **52**, 450–458 (1981).
3. Huybers, P. J. Early Pleistocene glacial cycles and the integrated summer insolation forcing. *Science* **313**, 508–511 (2006).
4. Huybers, P. J. & Tziperman, E. Integrated summer insolation forcing and 40,000 year glacial cycles: the perspective from an icesheet/energy balance model. *Paleoceanography* (in the press).
5. Clark, P. U. & Pollard, D. Origin of the middle Pleistocene transition by ice sheet erosion of regolith. *Paleoceanography* **13**, 1–9 (1998).
6. Raymo, M. E., Lisiecki, L. & Nisancioglu, K. Plio-Pleistocene ice volume, Antarctic climate, and the global  $\delta^{18}\text{O}$  record. *Science* **313**, 492–495 (2006).
7. Naish, T., Powell, R., Levy, R. & the ANDRILL-MIS Science Team. Examining Antarctica. *Geotimes* 30–33 (October 2007).
8. Lisiecki, L. E. & Raymo, M. E. A Plio-Pleistocene stack of 57 globally distributed benthic  $\delta^{18}\text{O}$  records. *Paleoceanography* **20**, PA1003 (2005).

**Acknowledgements** This work was supported by the National Science Foundation.

**Author Information** Reprints and permissions information is available at [npg.nature.com/reprints](http://npg.nature.com/reprints). Correspondence should be addressed to M.E.R. ([raymo@bu.edu](mailto:raymo@bu.edu)).

# Ocean circulation in a warming climate

J. R. Toggweiler & Joellen Russell

**Climate models predict that the ocean's circulation will weaken in response to global warming, but the warming at the end of the last ice age suggests a different outcome.**

There is an old truism in climate circles that the cold climate at the Last Glacial Maximum (LGM), which occurred 21,000 years ago, had stronger winds. This idea fits with the common observation that it is windier in the winter than in the summer because there is greater thermal contrast within the atmosphere in the winter hemisphere. Temperature reconstructions from the LGM show that Equator-to-pole gradients in sea surface temperature were indeed larger — that is, the polar oceans were colder than the tropical ocean at the LGM in comparison with the temperature differences today.

It is now becoming clear that the winds in the atmosphere drive most of the circulation in the ocean. If the LGM climate really did have stronger winds, it would thus be expected that the circulation in the ocean was more vigorous. The oceans seem to tell a different story, however. The deep water in the ocean's interior is continuously being replaced ('overturned') by surface waters from the poles. This overturning circulation in the Atlantic Ocean seems to have been weaker at the LGM<sup>1</sup>. The water in the deep ocean was also very 'old' in relation to the atmosphere — in terms of having a low radiocarbon content — indicating that the ocean's interior was poorly mixed and poorly ventilated<sup>2</sup>. The overturning circulation then seems to have strengthened as Earth began to warm about 18,000 years ago. The increased overturning vented the radiocarbon-depleted carbon dioxide (CO<sub>2</sub>) to the atmosphere, as seen in a pair of big dips in the radiocarbon activity of the atmosphere and upper ocean<sup>3</sup>. This addition of CO<sub>2</sub> to the atmosphere helped to warm the climate and bring the last ice age to an end.

These findings present a conundrum. If the winds were stronger in the cold glacial state and became weaker going into the warm interglacial state, then why was the ocean's circulation weaker during the cold glacial period? And how did it increase in strength during the transition to the warm interglacial period, causing the ocean's interior to become better mixed and better ventilated? Are researchers missing something about the factors that affect ocean circulation, or is it the old truism about the strength of the winds during the cold glacial period that is flawed?

During the 1990s, the first generation of coupled climate models predicted that the ocean's overturning circulation would weaken markedly over the next 100–200 years in response to global warming<sup>4</sup>. The predicted weakening is a response to the warming itself and to a stronger hydrological cycle, both of which make the ocean surface waters in the models less dense and less able to sink in relation to the water below. Thus, the models suggested that circulation would be less vigorous in a warming climate, somewhat like the weakening expected from diminished winds in a warmer climate outlined above. But again, the real ocean became better mixed and better ventilated when Earth began to warm about 18,000 years ago. So what will happen to the ocean's circulation in a warming climate? Are the models getting it wrong?

## Winds and the ocean's overturning circulation

Until recently, the circulation of the ocean was thought to comprise two fairly independent parts. The wind-driven circulation drove the surface currents in the ocean gyres, whereas the overturning circulation ventilated the interior with cold and relatively saline water from the

poles. The latter was called the 'thermohaline' circulation to emphasize that it was driven by buoyancy forces — warming, cooling, freshening and salinification — rather than the stress on the surface coming from the winds.

The inconsistencies mentioned earlier could be overlooked if this dichotomy holds, because the winds and the wind-driven circulation in the upper ocean could still have been stronger during the LGM while the thermohaline circulation was less vigorous. However, the dichotomy and the use of the term 'thermohaline' have almost disappeared from the oceanographic literature, because the circulation in the interior is now increasingly seen as being driven by turbulent mixing from the winds and tides<sup>5,6</sup> and directly by the winds themselves<sup>7</sup>.

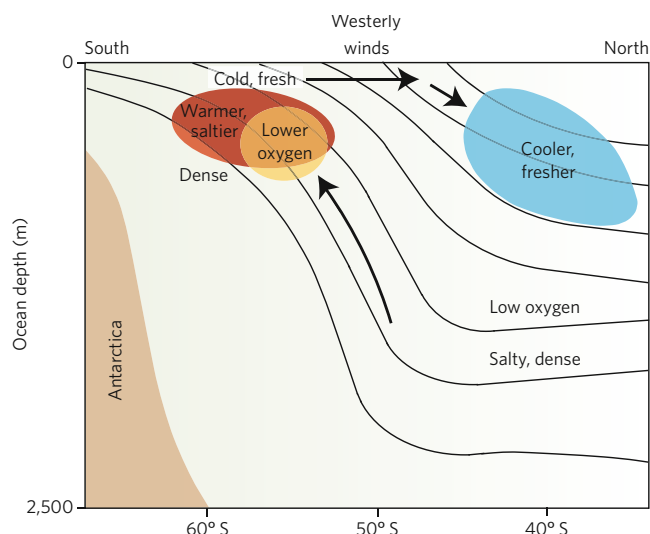
The westerly winds over the Southern Ocean seem to be crucial in this regard<sup>7</sup>. The Antarctic Circumpolar Current (ACC) is a wind-driven current that goes around Antarctica through an east–west channel between South America, Australia and Antarctica that is not blocked by land. Because the winds over the channel and the flow of the ACC are aligned for the length of the channel, the ACC is easily the world's strongest current (by volume of water transported). According to Carl Wunsch<sup>8</sup>, about 70% of the wind energy going into ocean currents globally goes directly into the ACC.

The same dense water found in the interior north of the ACC is also found just below the surface around Antarctica, and the westerly winds driving the ACC draw this dense water directly up to the surface (Fig. 1). In this way, the winds driving the ACC continually remove dense water from the interior. Dense water must sink elsewhere to replace the water drawn up by the winds around Antarctica.

The ACC is constrained to flow south of the tip of South America at 56° S as it passes from the Pacific Ocean to the Atlantic Ocean. Its mean position therefore lies between 50° S and 55° S. The strongest westerly winds tend to be found between 45° S and 50° S. This means that the strongest westerly winds are not actually aligned with the ACC. It is noteworthy in this regard that the westerly winds in both hemispheres have been shifting polewards and getting stronger over the past 40 years<sup>9,10</sup>, partly in response to the warming from higher atmospheric CO<sub>2</sub> concentrations<sup>11,12</sup>. Thus, the strongest westerlies are now more squarely over the ACC, and — as expected — they seem to be doing more work to drive the ACC and more work to draw deep water up to the surface than they were 40 years ago<sup>13,14</sup>.

Measurements south of Australia indicate that the ACC has strengthened since the 1960s and 1970s (ref. 15). The ocean's surface stands higher north of the ACC and lower south of the ACC than it did in the 1960s and 1970s, and changes in subsurface water properties show the pattern expected from a stronger wind effect (Fig. 1).

The first generation of climate models suggested that warmer ocean temperatures and the freshening of the polar oceans are the primary influences on the ocean's overturning circulation in a warming climate. Warmer ocean temperatures lead to more evaporation from the tropical ocean and more freshwater input to the polar oceans through precipitation and runoff from the land; that is, a stronger hydrological cycle. According to these models, polar freshening should have already led to



**Figure 1** Cross-section of the ACC, illustrating how stronger winds over the ACC lead to a stronger overturning circulation. The curved lines are isolines of constant density. These lines plunge downwards and to the north, reflecting the flow of the current (out of the page in the centre of the figure). Westerly winds above the ACC (also blowing out of the page) push cold, fresh surface waters away from Antarctica across the ACC (towards the blue area) and draw slightly warmer and salty water that is low in oxygen up from the interior to the surface (towards the red and yellow areas). Stronger winds in the past 40 years have resulted in more surface water being pushed northwards and have drawn more deep water up to the surface. As a result, the water just below the surface around Antarctica is now warmer, saltier and lower in oxygen, despite an overall freshening of the ocean around Antarctica. The water in the blue area to the north has become cooler and fresher. (Figure adapted from ref. 15.)

some weakening of the overturning and more stratification of the polar oceans in both hemispheres<sup>16</sup>. There is no firm evidence yet that this has happened<sup>17</sup>, possibly because stronger winds have maintained the circulation of salty water into the regions where sinking occurs. The first generation of climate models had weak winds and sluggish wind-driven circulations, and the winds in these models did not change with higher atmospheric CO<sub>2</sub>. Thus, the hydrological cycle in the early models had a free rein to slow the overturning as the climate warmed.

The climate models in the latest round of assessments by the Intergovernmental Panel on Climate Change predict that the westerlies will shift polewards and become stronger in the twenty-first century<sup>18</sup>. However, the models still suggest that the overturning of the Atlantic Ocean will weaken, although not nearly as much as in the first generation of climate models<sup>19</sup>.

### Role of temperature gradients

The poleward shift and the intensification of the westerlies over the past 40 years caught geoscientists by surprise. Because a higher atmospheric CO<sub>2</sub> concentration is supposed to warm the poles more than the tropics, the intensification, in particular, was not predicted. An important consideration in this regard is that the westerlies respond mainly to changes in the thermal contrast in the middle of the atmosphere rather than to changes at the surface, and the thermal contrast in the middle of the atmosphere has increased in response to higher CO<sub>2</sub> levels<sup>12</sup>.

A schematic illustration of the structure of the atmosphere is shown in Fig. 2, indicating how the atmosphere varies in response to higher CO<sub>2</sub> concentrations. CO<sub>2</sub> makes the atmosphere more opaque to outgoing long-wave radiation. More CO<sub>2</sub> warms the pocket of warm air near the surface in the tropics and subtropics, and cools the envelope of cold air above the tropics and subtropics and over the poles<sup>12</sup>. The position and strength of the westerlies reflect the thermal contrast between the pocket of warm air and the envelope of cold air.

At the LGM, a time of low atmospheric CO<sub>2</sub> concentrations, the pocket of warm air was cooler and probably did not extend as far above

the surface at low latitudes (Fig. 2b). This is consistent with the depression of the snowline seen on tropical mountains<sup>20</sup>. At the same time, the envelope of cold air should have been relatively warm. Thus, the thermal contrast in the middle of the atmosphere would have been relatively weak. From this perspective, weaker, not stronger, westerlies would be expected at the LGM.

The amount of CO<sub>2</sub> in the atmosphere increased at the end of the last ice age and is increasing again today. This increase seems to have warmed the pocket of warm air and caused it to expand upwards (Fig. 2a), whereas the envelope of cold air has cooled<sup>12</sup>. It is thought that the upward expansion and the cooling aloft have led to greater thermal contrast in the middle of the atmosphere, which has increased the strength of the mid-latitude westerlies and caused them to shift towards the poles. The wind stress on the ocean has become stronger and the position of maximum stress has shifted polewards in response to these changes in the westerly flow aloft.

At the LGM, by contrast, the strongest westerlies in the Southern Hemisphere seem to have been about 7–10° north of their modern position<sup>21</sup>. Because the ACC cannot change its position, a shift of this magnitude towards the Equator would have put the westerlies well to the north of the ACC, in a position where they could not put much energy into the ACC or the overturning circulation. A poleward shift and an intensification of the westerlies during the warming at the end of the last ice age would have put stronger westerlies closer to the ACC and might thus have enhanced the ocean's circulation, as postulated earlier.

### Ocean temperatures and the ozone hole

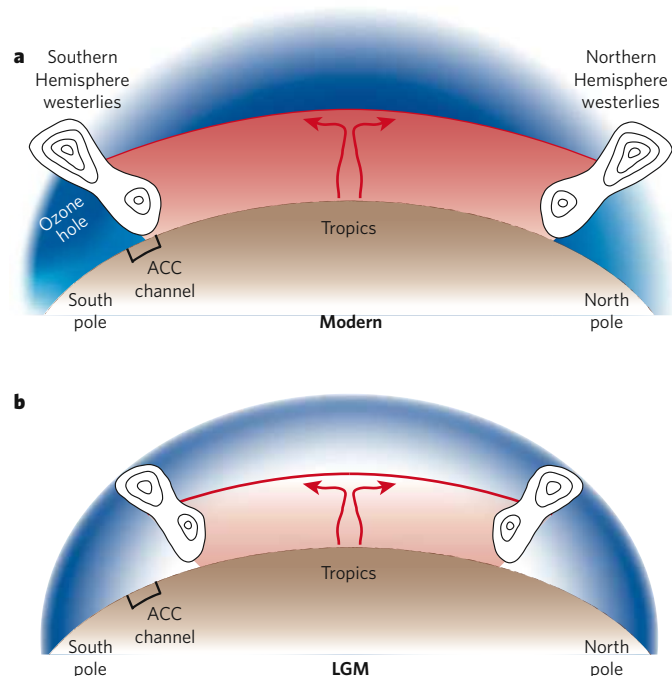
Two factors make the warming at the end of the last ice age and the warming today rather different. One is the rate of warming. The other factor is the ozone hole over Antarctica.

The temperature of the ocean is a crucial factor for the overturning, because the density of sea water responds more strongly to temperature when the ocean is warm and responds more strongly to salinity when the ocean is cold. The polar oceans were very cold at the peak of the last ice age. At these low temperatures, the overturning would have been very sensitive to inputs of fresh water near the poles<sup>22,23</sup>. Indeed, a cap of cold, fresh polar surface waters and extensive sea ice seems to have blocked the overturning around Antarctica and trapped a large quantity of radiocarbon-depleted CO<sub>2</sub> in the deep ocean<sup>24</sup>. As the ocean warmed and the westerlies shifted polewards, the cap seems to have broken down and released this CO<sub>2</sub> to the atmosphere<sup>21</sup>.

The warming and the release of CO<sub>2</sub> at the end of the last ice age were spread over several thousand years. At this pace, all the different parts of the ocean would presumably have warmed together. Most of the warming in the future, however, is expected to happen in the next 200 years. Thus, much of the future warming in the ocean could be confined to the ocean's surface layers. A surface-confined warming would work together with the hydrological cycle to weaken the overturning. However, if stronger winds can maintain the overturning, the warming in the future will be more evenly distributed through the ocean and not be as much of a factor.

The shift in the westerlies over the past 40 years has been asymmetrical, with a much larger shift in the south than in the north. The asymmetry is due at least partly to the depletion of stratospheric ozone over Antarctica, which was caused by the emission of long-lived chlorofluorocarbons during the twentieth century. Removing the ozone from the lower stratosphere is an effective way to cool the envelope of cold air over Antarctica (Fig. 2). Thus, the depletion of ozone, like the increase in CO<sub>2</sub> concentration, has increased the thermal contrast in the south and helped to make the southern westerlies stronger<sup>25</sup>.

The amount of ozone should be returning to previous levels over the next 40 years. This means that the wind effect on the ocean's overturning due to ozone depletion should be tailing off as the wind effect due to CO<sub>2</sub> continues to increase. Thus, the wind effect on the overturning might not be increasing as much over the next 40 years as it did over the past 40 years.



**Figure 2 | Changes in the westerlies and atmospheric structure in response to different CO<sub>2</sub> concentrations.** Bands of westerly winds in the Northern Hemisphere and Southern Hemisphere (shown schematically by the isotachs) separate the warm air (red shades) in the tropics from the cold air (blue shades) over the poles. **a**, Atmospheric structure today. Over recent decades, higher CO<sub>2</sub> concentrations have made the warm air warmer and the surrounding envelope of cold air cooler, especially near the top of the troposphere (curved red line). The thermal contrast across the zones of strong westerlies in the Northern and Southern Hemispheres is therefore greater, and the westerlies have become stronger and have shifted polewards in response. **b**, Proposed atmospheric structure at the LGM. With less CO<sub>2</sub> in the atmosphere, the thermal contrast in the middle of the atmosphere was probably decreased (indicated by paler shades), and the westerlies aloft should therefore have been relatively weak. The strongest westerlies were also significantly north of the ACC, where they would have had much less impact on the ocean.

## Lessons from the past

Anthropogenic additions of CO<sub>2</sub> to the atmosphere have resulted in a stronger hydrological cycle and a warming of the upper ocean that are currently threatening to weaken the ocean's overturning circulation. However, larger differences in temperature in the middle of the atmosphere have given rise to stronger winds that are acting to strengthen the circulation, as we argue they did at the end of the last ice age. What is uncertain is whether stronger winds and a stronger circulation will counter the freshening and distribute the extra heat through the interior over the next 200 years.

Current climate-system models say that the ocean's overturning circulation will weaken over the next century<sup>19</sup>, but these predictions might not rest on a solid foundation. The early climate models were deficient because they understated the effects of the winds in general and failed to anticipate the poleward shift and the intensification of the westerlies over the past 40 years. The latest models are much improved but might still not fully represent the wind effect.

A key test for the models is to reproduce the changes that took place at the end of the last ice age. Does the oceanic circulation in the models get weak enough in a cold LGM-like state to bottle up so much CO<sub>2</sub>? More importantly, can the weaker circulation make the CO<sub>2</sub> in the deep ocean very old with respect to the radiocarbon activity in the atmosphere<sup>22</sup>? Can

the circulation then get strong enough to let all the radiocarbon-depleted CO<sub>2</sub> back out? From the observations, it is clear that large circulation changes took place, and it seems unlikely that circulation changes of this magnitude could have happened without substantial changes in the wind forcing. It seems that the information from the past is telling us to expect a stronger oceanic circulation in the warmer climate to come.

J. R. Toggweiler is at the Geophysical Fluid Dynamics Laboratory, National Oceanic and Atmospheric Administration, Princeton, New Jersey 08542, USA. Joellen Russell is in the Department of Geosciences, University of Arizona, Tucson, Arizona 85721, USA.

- Lynch-Stieglitz, J. *et al.* Atlantic meridional overturning during the Last Glacial Maximum. *Science* **316**, 66–69 (2007).
- Sikes, E. L., Samson, C. R., Guilderson, T. P. & Howard, W. R. Old radiocarbon ages in the Southwest Pacific at 11,900 years ago and the last glaciations. *Nature* **405**, 555–559 (2000).
- Marchitto, T. M., Lehman, S. J., Ortiz, J. D., Fluckiger, J. & van Geen, A. Marine radiocarbon evidence for the mechanism of deglacial atmospheric CO<sub>2</sub> rise. *Science* **316**, 1456–1459 (2007).
- Manabe, S. & Stouffer, R. J. Century-scale effects of increased atmospheric CO<sub>2</sub> on the ocean-atmosphere system. *Nature* **364**, 215–218 (1993).
- Munk, W. & Wunsch, C. Abyssal recipes II: energetics of tidal and wind mixing. *Deep-Sea Res.* **45**, 1977–2010 (1998).
- Kuhlbrodt, T. *et al.* On the driving processes of the Atlantic meridional overturning circulation. *Rev. Geophys.* **45**, RG2001, doi:10.1029/2004RG000166 (2007).
- Toggweiler, J. R. & Samuels, B. Effect of Drake Passage on the global thermohaline circulation. *Deep-Sea Res.* **42**, 477–500 (1995).
- Wunsch, C. Work done by the wind on the oceanic general circulation. *J. Phys. Oceanogr.* **28**, 2332–2340 (1998).
- Hurrell, J. W. & van Loon, H. A modulation of the atmospheric annual cycle in the Southern Hemisphere. *Tellus A* **46**, 325–338 (1994).
- McGabe, G. J., Clark, M. P. & Serreze, J. C. Trends in Northern Hemisphere surface cyclone frequency and intensity. *J. Clim.* **14**, 2763–2768 (2001).
- Gillett, N. P., Zwiers, F. W., Weaver, A. J. & Stott, P. A. Detection of human influence on sea level pressure. *Nature* **422**, 292–294 (2003).
- Shindell, D. T. & Schmidt, G. A. Southern Hemisphere climate response to ozone changes and greenhouse gas increases. *Geophys. Res. Lett.* **31**, L18209, doi:10.1029/2004GL020724 (2004).
- Saenko, O. A., Fyfe, J. C. & England, M. H. On the response of the ocean wind-driven circulation to atmospheric CO<sub>2</sub> increase. *Clim. Dyn.* **25**, 415–426 (2005).
- Russell, J. L., Dixon, K. W., Gnanadesikan, A., Stouffer, R. J. & Toggweiler, J. R. The Southern Hemisphere westerlies in a warming world: Propping open the door to the deep ocean. *J. Clim.* **19**, 6382–6390 (2006).
- Aoki, S., Bindoff, N. L. & Church, J. A. Interdecadal water mass changes in the Southern Ocean between 30°E and 160°E. *Geophys. Res. Lett.* **32**, L07607, doi:10.1029/2004GL022220 (2005).
- Sarmiento, J. L., Hughes, T. M. C., Stouffer, R. J. & Manabe, S. Simulated response of the ocean carbon cycle to anthropogenic climate warming. *Nature* **393**, 245–249 (1998).
- Bindoff, N. L. *et al.* in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Solomon, S. *et al.*) 385–432 (Cambridge Univ. Press, Cambridge, UK, 2007).
- Yin, J. H. A consistent poleward shift of the storm tracks in simulations of 21st century climate. *Geophys. Res. Lett.* **32**, L18701, doi:10.1029/2005GL023684 (2005).
- Gregory, J. M. *et al.* A model intercomparison of changes in the Atlantic thermo-haline circulation in response to increasing atmospheric CO<sub>2</sub> concentration. *Geophys. Res. Lett.* **32**, L12703, doi:10.1029/2005GL023209 (2005).
- Broecker, W. S. & Denton, G. H. The role of ocean-atmosphere reorganizations in glacial cycles. *Geochim. Cosmochim. Acta* **53**, 2465–2501 (1989).
- Toggweiler, J. R., Russell, J. L. & Carson, S. R. Midlatitude westerlies, atmospheric CO<sub>2</sub>, and climate change during the ice ages. *Paleoceanography* **21**, PA2005, doi:10.1029/2005PA001154 (2006).
- Sigman, D. M., Jaccard, S. L. & Haug, G. H. Polar ocean stratification in a cold climate. *Nature* **428**, 59–63 (2004).
- De Boer, A. M., Sigman, D. M., Toggweiler, J. R. & Russell, J. L. Effect of global ocean temperature changed on deep ocean ventilation. *Paleoceanography* **22**, PA2210, doi:10.1029/2005PA001242 (2007).
- Francois, R. F. *et al.* Water column stratification in the Southern Ocean contributed to the lowering of glacial atmospheric CO<sub>2</sub>. *Nature* **389**, 929–935 (1997).
- Thompson, D. W. J. & Solomon, S. Interpretation of recent Southern Hemisphere climate change. *Science* **296**, 895–899 (2002).

**Acknowledgements** We thank I. Held and M. Wallace for critical insights, A. Gnanadesikan for comments on the manuscript, and C. Raphael and J. Varanyak for help with the figures. J.R.'s work was supported by a grant from the National Oceanic and Atmospheric Administration.

**Author Information** Reprints and permissions information is available at [npg.nature.com/reprints](http://npg.nature.com/reprints). Correspondence should be addressed to J.R.T. ([robbie.toggweiler@noaa.gov](mailto:robbie.toggweiler@noaa.gov)).

# Terrestrial ecosystem carbon dynamics and climate feedbacks

Martin Heimann & Markus Reichstein

**Recent evidence suggests that, on a global scale, terrestrial ecosystems will provide a positive feedback in a warming world, albeit of uncertain magnitude.**

It has only been recognized relatively recently that biological processes can control and steer the Earth system in a globally significant way. Terrestrial ecosystems constitute a major player in this respect: they can release or absorb globally relevant greenhouse gases such as carbon dioxide ( $\text{CO}_2$ ), methane and nitrous oxide, they emit aerosols and aerosol precursors, and they control exchanges of energy, water and momentum between the atmosphere and the land surface. Ecosystems themselves are subject to local climatic conditions, implying a multitude of climate–ecosystem feedbacks that might amplify or dampen regional and global climate change. Of these feedbacks, that between the carbon cycle and climate has recently received much attention. Large quantities of carbon are stored in living vegetation and soil organic matter, and liberation of this carbon into the atmosphere as  $\text{CO}_2$  or methane would have a serious impact on global climate. By definition, the carbon balance of an ecosystem at any point in time is the difference between its carbon gains and losses. Terrestrial ecosystems gain carbon through photosynthesis and lose it primarily as  $\text{CO}_2$  through respiration in autotrophs (plants and photosynthetic bacteria) and heterotrophs (fungi, animals and some bacteria), although losses of carbon as volatile organic compounds, methane or dissolved carbon (that is, non- $\text{CO}_2$  losses) could also be significant. Quantifying and predicting these carbon-cycle–climate feedbacks is difficult, however, because of the limited understanding of the processes by which carbon and associated nutrients are transformed or recycled within ecosystems, in particular within soils, and exchanged with the overlying atmosphere.

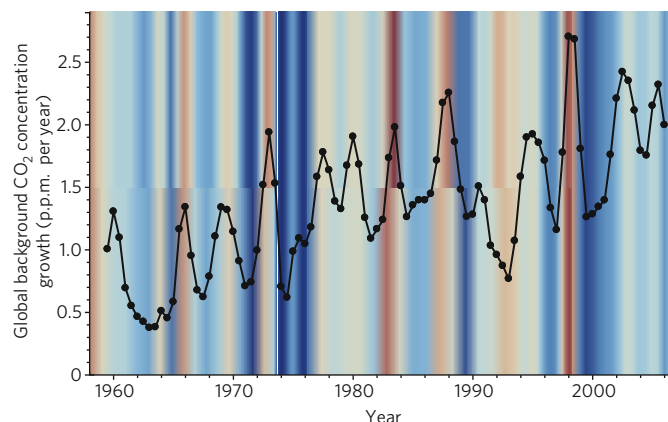
There is ample empirical evidence that the terrestrial component of the carbon cycle is responding to climate variations and trends on a global scale. This is exemplified by the strong interannual variations in the globally averaged growth rate of atmospheric  $\text{CO}_2$ , which is tightly correlated with El Niño–Southern Oscillation climate variations (Fig. 1). Many lines of evidence show that the variations in the  $\text{CO}_2$  growth rate are mainly caused by terrestrial effects, in particular the impacts of heat and drought on the vegetation of western Amazonia and southeastern Asia, leading to ecosystem carbon losses through decreased vegetation productivity and/or increased respiration. These interannual variations reflect short-term responses of the carbon cycle to climate perturbations, however, and cannot be expected to hold over longer timescales. Conversely, the close correlation between atmospheric concentrations of  $\text{CO}_2$ , methane and nitrous oxide and global climate during the last glacial cycles<sup>1</sup> indicates that ecosystem–climate interactions are also operating on timescales of millennia and longer.

Unfortunately, empirical evidence for global carbon-cycle–climate interactions on the timescale pertinent to current global climate change, that is, decades to centuries, is much scarcer. Hence the assessment on these timescales has to be attempted by means of comprehensive, coupled carbon-cycle–climate models. A recent comparison of different model simulations for the industrial epoch (the past ~150 years) and the next 100 years, made on the basis of a standard model of  $\text{CO}_2$  emissions,

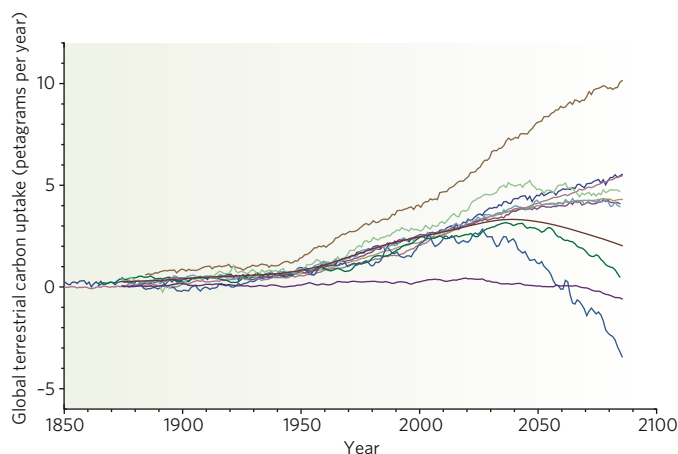
has shown a variety of responses<sup>2</sup>. Almost all the models show terrestrial  $\text{CO}_2$  sequestration in the early phase of industrial expansion in the nineteenth and twentieth centuries but a substantial decrease in sequestration as the world warms (Fig. 2) (see page 297). In some models, the terrestrial carbon cycle even becomes a substantial source of atmospheric  $\text{CO}_2$  and thus strongly amplifies global climate change. The rather wide spread of results from the different model simulations demonstrates on the one hand genuine differences in the simulated climate change, and on the other hand the very poor understanding of processes in functioning ecosystems as represented in these models.

## Changing concepts of ecosystem carbon dynamics

In carbon-cycle–climate models, the effect of the prevailing climate on the carbon balance in terrestrial ecosystems is described mostly by relatively simple response functions and kinetic concepts of  $\text{CO}_2$  uptake by photosynthesis and loss by respiration. The fundamental paradigm adopted by researchers over the past two decades has been that photosynthetic uptake is stimulated both by increasing  $\text{CO}_2$  and, in boreal and temperate regions, by rising temperature, although both effects are expected to saturate at high levels of these variables. On the other hand, the biological processes underlying respiration are assumed to respond to temperature in an exponential way but are not affected by the  $\text{CO}_2$  concentration<sup>3,4</sup>. This leads to the conclusion that the biosphere is able to provide negative feedback to rising  $\text{CO}_2$  and temperature until the temperature climbs so high that the stimulating



**Figure 1 | Estimated growth rate of the global background atmospheric  $\text{CO}_2$  concentration.** Global  $\text{CO}_2$  concentration is estimated from measurements from the South Pole and the Mauna Loa (Hawaii) long-term monitoring stations (ref. 17, updated). The black dots represent centred annual averages calculated at six-monthly intervals. The coloured background shows the variation of the multivariate El Niño–Southern Oscillation index. Blue shades indicate negative phases, and brown shades positive phases, of this index<sup>18</sup>. p.p.m., parts per million.



**Figure 2 | Comparison of estimated global terrestrial carbon uptake in different models of the carbon-cycle-climate system.** Global terrestrial carbon uptake was simulated by 11 coupled carbon-cycle-climate models driven with carbon emissions from the SRES-A2 emissions profile. Data are taken from the Coupled Carbon Cycle Climate Model Intercomparison Project<sup>2</sup>, with uptake rates smoothed with a 30-year moving average.

effect on respiration exceeds the CO<sub>2</sub> fertilization effect. This fundamental principle reflects the behaviour of almost all the models in the comparative study described earlier<sup>2</sup>.

The fundamental simplifying assumption behind this reasoning is that above-ground assimilatory processes (plant photosynthesis) and below-ground heterotrophic respiratory processes (for example, decomposition by fungi and respiration by animal and bacterial life in the soil) can be conceptually isolated and analysed separately. Although this conceptual model has provided valuable guidance for experimental and model design, evidence has accumulated in recent years that above- and below-ground processes are intimately linked, constituting a complex and dynamic system with non-negligible interactions. Hence, the situation is much more complicated than previously thought and might result in unexpected dynamics through interactions between physical, chemical and biological processes within the ecosystem — particularly in the soil. This implies that, beyond rising CO<sub>2</sub> levels and rising temperature, other climatic and environmental factors might modify, or even dominate, the carbon balance of the world's ecosystems. Furthermore, not only the long-term rate of change of mean values of parameters such as temperature but also alterations in their variability, including greater extremes, may be crucial to ecosystem carbon dynamics.

### Ecosystems in a multi-factor world

Primary productivity in more than half of the world's ecosystems is substantially limited by the availability of water. Hence, changes in precipitation will have direct effects on ecosystem carbon dynamics. In a warmer world, evaporation is expected to increase, leading to a more negative water balance, whereas decreased water loss through stomata in a CO<sub>2</sub>-richer world will tend to mitigate this effect. The net effect (production minus respiration) of a more negative overall water balance probably depends on the water-holding capacity of the soil, the vertical distribution of carbon and roots in the soil, and the general drought sensitivity of the vegetation. For instance, if most of the soil carbon is concentrated at the top of the soil, while roots go deep into a soil with high water-holding capacity, or even tap the groundwater, soil carbon decomposition will initially be more strongly affected by drought than will vegetation productivity, as the topsoil dries out first. Water limitation may even suppress the effective ecosystem-level response of temperature on respiration<sup>5</sup>. Conversely, if soil water-holding capacity is low, as in shallow soils, vegetation productivity will be strongly affected by a negative water balance. Hence, under drier conditions, there are predictions of increased sequestration by suppression of respiration and of net loss of carbon through decreased productivity<sup>6,7</sup>.

A second important interacting factor is the available nitrogen, which often determines the magnitude of the CO<sub>2</sub> fertilization effect and may suppress it completely if nitrogen is limiting<sup>8,9</sup> (see page 293). There are also indications of strong interactions between water and nitrogen, with nitrogen becoming more limiting under drier conditions. Other factors to be considered are changes in the amount and quality (direct or diffuse) of light, which can alter vegetation productivity<sup>10</sup>, and increases in air pollutants and ozone, with their detrimental effects on primary production<sup>11</sup>.

### Climate variability and extremes

The terrestrial biosphere does not respond to a mean climate but to the concrete time series of actual weather conditions. Consequently, anticipated reactions to gradual mean changes in climate components and atmospheric concentrations of trace gases might be misleading if variability and extremes are not considered. A recent wake-up call in this respect was the European heatwave in the summer of 2003, when the cumulative European carbon sequestration of five years was undone within a few months through the reaction of the terrestrial biosphere to these extreme hydrological and climatic conditions<sup>6</sup>. Lag effects — for example, increased tree death in the years after an extreme event — may yet increase the effect of the heatwave on European ecosystems. Apart from extremes, changes in the seasonal distribution of climate factors may be decisive. This is particularly evident for water-carbon-cycle interactions, where changes in the frequency or timing of rainfall without changes in the annual total may have profound effects on ecosystem productivity<sup>12</sup>, as these factors determine whether the water will be used by plants and transpired, or will just run off or evaporate.

Similarly, temporal changes in constellations of water deficit, wind speed, air temperature and humidity modify the frequency and severity of forest fires and the consequent rapid loss of carbon from the biosphere. Wind-throws due to a single large storm kill trees and so make previously 'locked-in carbon' subject to decay and release of CO<sub>2</sub>. Changes in the seasonality of temperature can also have consequences; for example, the warmer winter and spring in large parts of the Northern Hemisphere in 2006/2007 induced earlier leafing and flowering, leading to greater vulnerability of plants to late frosts. Our predictive ability in respect of such local weather conditions is clearly limited by both the level of detail that can be incorporated into atmosphere-ocean general circulation models and our understanding of the seasonal dynamics of ecosystems and their ability to acclimate on a variety of timescales.

### Nonlinear ecosystem feedback loops

As discussed above, the net effect of any environmental change on the carbon balance in an ecosystem depends on the reactions of both photosynthesis and respiration; in other words, on above-ground and below-ground processes. Below-ground processes in particular are still poorly understood yet provide a number of potentially important feedbacks in the carbon-cycle-climate system. Here, we focus on below-ground processes and recent important findings on biological-physicochemical interactions that are not considered in current simulations of the carbon-cycle-climate system; Figure 3 illustrates three exemplary and simplified conceptual descriptions of subsystems by means of cause-and-effect pathways that are related to the dynamics of ecosystem carbon.

Figure 3a shows potential interactions between microbial metabolism and the physics of permafrost thawing and carbon release. Current estimates of carbon stored deep-frozen in permafrost regions amount to at least 400 petagrams ( $4 \times 10^{11}$  tonnes) of carbon (ref. 13) that is relatively unprocessed and labile as the frozen state protects it from microbial decomposition. Moss and turf layers provide very good insulation against the atmosphere. With rising summer temperatures, these soils begin to melt, the carbon becomes metabolized and microbial metabolism may release enough heat (the 'dung-heap effect') to facilitate further melting, providing a nonlinear positive-feedback mechanism to enhance permafrost melting and, through methane and CO<sub>2</sub> emissions, to increase the greenhouse effect. Model simulations indicate that a run-away dynamic

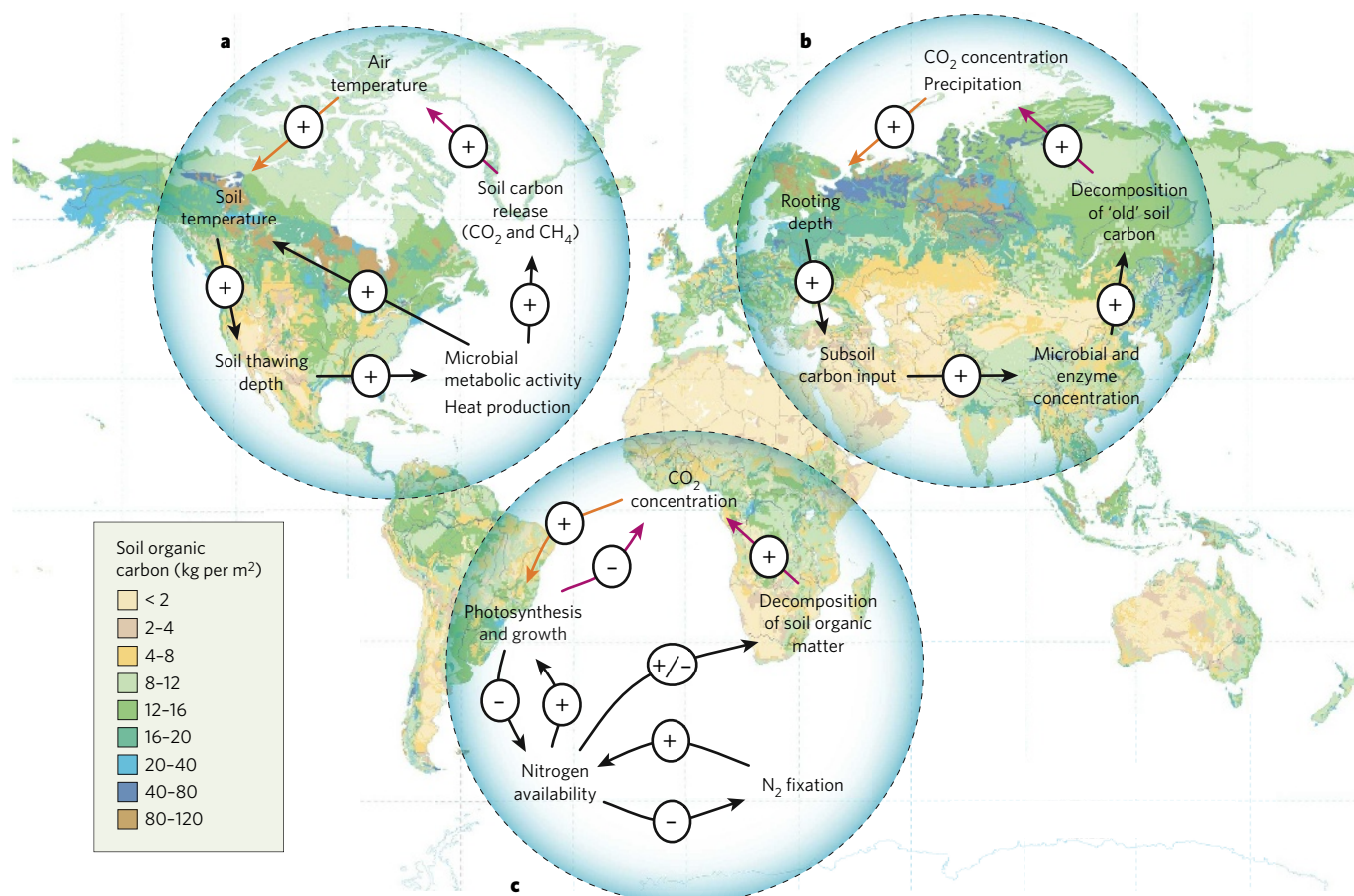
may be triggered by a few warm years, but the strength of this feedback mechanism and the realism of these simulations remain unclear<sup>14</sup>.

Another mechanism for potential mobilization of large amounts of carbon is the so-called 'microbial priming effect'. It has been shown in several experimental systems that the addition of substrates with readily available energy (for example, glucose and cellulose) to the soil stimulates the decomposition of 'old' soil carbon. Sébastien Fontaine *et al.*<sup>15,16</sup> showed that simply by adding cellulose to the soil they could mobilize carbon from the subsoil of grasslands that was assumed to be stable, whereas other factors such as temperature, nitrogen addition or increasing oxygen concentration had no effect. Counterintuitively, addition of such material even induced a net loss of carbon from the soil samples, as the soil carbon stock is large. In the context of climate change this effect may induce a positive-feedback effect, particularly in grassland soils (Fig. 3b). Increasing CO<sub>2</sub> concentrations can lead to enhanced below-ground allocation of labile carbon through roots and root exudates, which can enhance microbial activity and foster decomposition of carbon material that has been deemed stable but was in fact not being attacked because microbes were not active. Also, if rooting patterns change, either because of altered precipitation or as part of general vegetation dynamics, carbon input into deeper layers that were not rooted before might induce release of old carbon through this mechanism.

Last but not least, the interaction of the carbon and nitrogen cycles offers a plethora of mechanisms that could alter expected ecosystem carbon responses to the prevailing trend in climate change. Some of these are shown in Fig. 3c. In nitrogen-limited ecosystems, nitrogen

nutrition limiting the CO<sub>2</sub> fertilization effect on canopy assimilation is regularly found after a few years of increasing CO<sub>2</sub> levels<sup>9</sup>. There are also indications that nitrogen availability influences the decomposition of soil organic matter. Fungi use lignin, an abundant, stable organic substance found in plant cell walls, as a nitrogen source under conditions of limited nitrogen availability. Enhanced decomposition of lignin may lead to a positive feedback in response to rising atmospheric CO<sub>2</sub>. On timescales longer than a few years, however, acclimation or change in species composition, or, for example, increased nitrogen fixation through increased carbohydrate input into the soil, may relax or even overcompensate for the nitrogen-limitation effects. Also, an interaction with microbial 'priming' (see above) through more intensive and deeper plant rooting is not unlikely, as a decrease in nitrogen availability often leads to a larger allocation of carbon to roots.

Thus, the picture of a gradual increase in CO<sub>2</sub> and temperature, with separable, non-interactive effects on assimilation and respiration, needs to be replaced by a multifactor view, by more sophisticated characterization of changes in environmental factors, including their variability and extremes, and, maybe most importantly, by stronger integrative consideration of complex interactions between ecosystem processes at different levels of organization. Most of these emerging characteristics point to a lower CO<sub>2</sub>-sequestration potential than estimated by current models and highlight the vulnerability of soil carbon that has accumulated over millennia. A positive feedback of ecosystem carbon to climate change might occur earlier and more strongly than currently predicted in coupled carbon-cycle-climate models<sup>2</sup>.



**Figure 3 | Feedback loops that could be induced by climate change in below-ground ecosystem carbon balances.** The three examples given here are crucial processes in the ecosystem, shown in simplified form. **a**, Potential interactions between microbial metabolism and the physics of permafrost thawing and carbon release. **b**, The 'microbial priming effect'. An increase in carbon and energy sources easily utilized by microbes can stimulate the decomposition of 'old' soil carbon, especially in grassland soils. In the context of climate change this effect may have a positive-feedback effect

on CO<sub>2</sub> increase and global warming. **c**, Interactions between the carbon and nitrogen cycles shown here could alter expected ecosystem carbon responses to the prevailing trend of climate change. Pink arrows denote effects of terrestrial ecosystems on climate, orange arrows denote effects of climate change on terrestrial ecosystems, and black arrows denote interactions within ecosystems. The background image is a world map of soil organic carbon. (Map reproduced, with permission, from USDA-NRCS, <http://soils.usda.gov/use/worldsoils/mapindex/soc.html>.)

### Future directions

It is evident that large uncertainties remain in our ability to assess terrestrial carbon-cycle–climate feedbacks over the coming decades. Current experiments give ambiguous results and do not provide definite conclusions on the importance of the mechanisms discussed above. Overall, it is likely that, at least on a global scale, terrestrial ecosystems will provide a positive, amplifying feedback in a warming world, albeit of uncertain magnitude. An important improvement in our understanding might be obtained by the combination of long-term multifactorial experiments with non-destructive ecosystem-level observations, such as whole-ecosystem flux measurements, and the integration of the results with ecosystem modelling in a multiple-constraint framework. As long as there is no fundamental understanding of the processes involved, simulations of coupled carbon-cycle–climate models can only illustrate the importance of, but do not show, a conclusive picture of the multitude of possible carbon-cycle–climate system feedbacks. Moreover, strong interactions between the natural processes described here and anthropogenic changes in land use, cover and management have to be expected. ■

Martin Heimann and Markus Reichstein are the Max Planck Institute for Biogeochemistry, Hans-Knöll-Strasse 10, D-07745 Jena, Germany.

- Petit, J. R. *et al.* Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399**, 429–436 (1999).
- Friedlingstein, P. *et al.* Climate–carbon cycle feedback analysis: results from the (CMIP)-M-4 model intercomparison. *J. Climate* **19**, 3337–3353 (2006).
- Kirschbaum, M. U. F. The temperature dependence of organic-matter decomposition — still a topic of debate. *Soil Biol. Biochem.* **38**, 2510–2518 (2006).
- Davidson, E. A. & Janssens, I. A. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* **440**, 165–173 (2006).
- Reichstein, M. *et al.* Determinants of terrestrial ecosystem carbon balance inferred from European eddy covariance flux sites. *Geophys. Res. Lett.* **34**, L01402, doi:10.1029/2006GL027880 (2007).
- Ciais, P. *et al.* Europe-wide reduction in primary productivity caused by the heat and drought in 2003. *Nature* **437**, 529–533 (2005).
- Saleska, S. R. *et al.* Carbon in Amazon forests: unexpected seasonal fluxes and disturbance-induced losses. *Science* **302**, 1554–1557 (2003).
- Hyvonen, R. *et al.* The likely impact of elevated [CO<sub>2</sub>], nitrogen deposition, increased temperature and management on carbon sequestration in temperate and boreal forest ecosystems: a literature review. *New Phytol.* **173**, 463–480 (2007).
- Reich, P. B. *et al.* Nitrogen limitation constrains sustainability of ecosystem response to CO<sub>2</sub>. *Nature* **440**, 922–925 (2006).
- Farquhar, G. D. & Roderick, M. L. Atmospheric science: Pinatubo, diffuse light, and the carbon cycle. *Science* **299**, 1997–1998 (2003).
- Sitch, S., Cox, P. M., Collins, W. J. & Huntingford, C. Indirect radiative forcing of climate change through ozone effects on the land–carbon sink. *Nature* **448**, 791–794 (2007).
- Knapp, A. K. *et al.* Rainfall variability, carbon cycling, and plant species diversity in a mesic grassland. *Science* **298**, 2202–2205 (2002).
- Sabine, C. L. *et al.* in *The Global Carbon Cycle: Integrating Humans, Climate and the Natural World* (eds Field, C. & Raupach, M.) 17–44 (Island, Washington DC, 2004).
- Khvorostyanov, D. V., Krinner, G., Ciais, P., Heimann, M. & Zimov, S. A. Vulnerability of permafrost carbon to global warming. Part 1. Model description and role of heat generated by organic matter decomposition. *Tellus* (in the press).
- Fontaine, S., Bardoux, G., Abbadie, L. & Mariotti, A. Carbon input to soil may decrease soil carbon content. *Ecol. Lett.* **7**, 314–320 (2004).
- Fontaine, S. *et al.* Stability of organic carbon in deep soil layers controlled by fresh carbon supply. *Nature* **450**, 277–280 (2007).
- Keeling, C. D. *et al.* Exchanges of Atmospheric CO<sub>2</sub> and <sup>13</sup>CO<sub>2</sub> with the Terrestrial Biosphere and Oceans from 1978 to 2000. I. Global aspects (Scripps Institution of Oceanography, San Diego, 2001).
- Wolter, K. & Timlin, M. S. Measuring the strength of ENSO events — how does 1997/98 rank? *Weather* **53**, 315–324 (1998).

**Author Information** Reprints and permissions information is available at [npg.nature.com/reprints](http://npg.nature.com/reprints). Correspondence should be addressed to M.H. ([martin.heimann@bgc-jena.mpg.de](mailto:martin.heimann@bgc-jena.mpg.de)).

# An Earth-system perspective of the global nitrogen cycle

Nicolas Gruber & James N. Galloway

**With humans having an increasing impact on the planet, the interactions between the nitrogen cycle, the carbon cycle and climate are expected to become an increasingly important determinant of the Earth system.**

The massive acceleration of the nitrogen cycle as a result of the production and industrial use of artificial nitrogen fertilizers worldwide has enabled humankind to greatly increase food production, but it has also led to a host of environmental problems, ranging from eutrophication of terrestrial and aquatic systems to global acidification. The findings of many national and international research programmes investigating the manifold consequences of human alteration of the nitrogen cycle have led to a much improved understanding of the scope of the anthropogenic nitrogen problem and possible strategies for managing it. Considerably less emphasis has been placed on the study of the interactions of nitrogen with the other major biogeochemical cycles, particularly that of carbon, and how these cycles interact with the climate system in the presence of the ever-increasing human intervention in the Earth system<sup>1</sup>. With the release of carbon dioxide (CO<sub>2</sub>) from the burning of fossil fuels pushing the climate system into uncharted territory<sup>2</sup>, which has major consequences for the functioning of the global carbon cycle, and with nitrogen having a crucial role in controlling key aspects of this cycle, questions about the nature and importance of nitrogen–carbon–climate interactions are becoming increasingly pressing. The central question is how the availability of nitrogen will affect the capacity of Earth's biosphere to continue absorbing carbon from the atmosphere (see page 289), and hence continue to help in mitigating climate change. Addressing this and other open issues with regard to nitrogen–carbon–climate interactions requires an Earth-system perspective that investigates the dynamics of the nitrogen cycle in the context of a changing carbon cycle, a changing climate and changes in human actions.

## The anthropogenic perturbation of the nitrogen cycle

Nitrogen is a fundamental component of living organisms; it is also in short supply in forms that can be assimilated by plants in both marine and land ecosystems. As a result, nitrogen has a critical role in controlling primary production in the biosphere. Nitrogen is also a limiting factor for the plants grown by humans for food. Without the availability of nitrogenous fertilizer produced by the industrial process known as the Haber–Bosch process, the enormous increase in food production over the past century, which in turn has sustained the increase in global population, would not have been possible. All the nitrogen used in food production is added to the environment, as is the nitrogen emitted to the atmosphere during fossil-fuel combustion. In the 1990s, these two sources of anthropogenic nitrogen to the environment amounted to more than 160 teragrams (Tg) N per year (Fig. 1). On a global basis, this is more than that supplied by natural biological nitrogen fixation on land (110 Tg N per year) or in the ocean (140 Tg N per year) (Fig. 1). Given expected trends in population, demand for food, agricultural practices and energy use, anthropogenic nitrogen fluxes are fated to increase; that is, humans are likely to be responsible for doubling the

turnover rates not only of the terrestrial nitrogen cycle but also of the nitrogen cycle of the entire Earth.

The negative consequences of these nitrogen additions are substantial and manifold, ranging from eutrophication of terrestrial and aquatic systems to global acidification and stratospheric ozone loss<sup>3</sup>. Of particular concern is the fact that chemical transformations of nitrogen along its transport pathway in the environment often lead to a cascade of effects. For example, an emitted molecule of nitrogen oxide can first cause photochemical smog and then, after it has been oxidized in the atmosphere to nitric acid and deposited on the ground, can lead to ecosystem acidification and eutrophication. Although there is still much to understand about the implications of nitrogen accumulation in the environment, there is also much to understand about how the increased availability of nitrogen interacts with other biogeochemical element cycles and how those interactions affect global climate change.

## Nitrogen and the perturbation of other element cycles

The human acceleration of the nitrogen cycle did not occur in isolation, as humans have altered the cycles of many other elements as well, most notably those of phosphorus, sulphur and carbon<sup>1</sup>. Of particular relevance is the acceleration of the global carbon cycle, because of the central role of atmospheric CO<sub>2</sub> in controlling climate<sup>4</sup>. As a result of the burning of fossil fuels and carbon emissions from land-use change, atmospheric CO<sub>2</sub> has increased to levels that are more than 30% above those of pre-industrial times. This increase in atmospheric CO<sub>2</sub> has been identified as the primary cause for the observed warming over the past century, particularly that of the past 30 years<sup>2</sup>.

The perturbations of the global nitrogen and carbon cycles caused by human activity are in part linked to each other. This is mostly a result of the atmosphere's being very efficient in spreading the nitrogen oxides and ammonia emitted as a result of energy and food production, and also because this nitrogen is deposited on the ground in a form that is readily available to plants, thereby stimulating productivity and enhancing the uptake of CO<sub>2</sub> from the atmosphere.

The existence of a largely unexplained, but substantial, carbon sink in the Northern Hemisphere terrestrial biosphere<sup>5</sup> (that is, in exactly the region that receives most of the anthropogenic nitrogen from the atmosphere) would seem to support this conjecture. However, nitrogen-addition and modelling studies suggest that the contribution of nitrogen fertilization to the Northern Hemisphere carbon land sink has been small. This issue needs to be resolved, because the different processes that are being considered to explain the current Northern Hemisphere carbon sink have very different future trajectories. If CO<sub>2</sub> fertilization is responsible — that is, the direct effect of elevated CO<sub>2</sub> on plant growth — one could expect this process to continue largely unabated into the future. If nitrogen fertilization is responsible, however, one could expect

the effect to level off in the future, primarily because the effect tends to decrease with increasing nitrogen load<sup>6</sup>.

The deposition of biologically available nitrogen into the ocean could also fertilize the ocean's biosphere and stimulate additional uptake of CO<sub>2</sub> there. On a global scale, the atmospheric deposition is small relative to the amount of nitrogen that is being fixed into organic matter and exported to depth, but it is an important source of external reactive nitrogen, being second in importance to naturally occurring marine nitrogen fixation (Fig. 1). The relative contribution of atmospherically derived reactive nitrogen to the total nitrogen demand can be much larger in certain regions, particularly in coastal regions downwind of the major Northern Hemisphere sources, and in regions where the vertical supply of reactive nitrogen from below is very restricted, such as the central subtropical ocean gyres.

The coastal ocean also receives a significant amount of anthropogenic nitrogen through rivers (Fig. 1). In some areas, this has led to well documented coastal eutrophication<sup>7</sup>, but the general consensus has been that the anthropogenic increase in river-derived nitrogen has had no impact on the open ocean.

### Elemental interactions of the natural cycles

The natural (unperturbed) components of the carbon and nitrogen cycles are even more tightly coupled than are the anthropogenic components (Fig. 1). This is a direct consequence of the presence of life, which links the elemental cycles of carbon, nitrogen and other elements at the molecular level, as a result of the constitutional need of organisms for these elements to build their tissues. This coupling occurs with specific elemental stoichiometries, whose values and flexibilities determine not only the relative speed at which the different cycles are coupled, but also how tight the coupling is<sup>8</sup>. In the ocean, the C/N ratio of the autotrophic phytoplankton responsible for nearly all marine photosynthesis varies remarkably little, whereas the C/N ratio of terrestrial plants is substantially more variable and also tends to be larger than that for marine phytoplankton.

Understanding the processes that control the C/N ratios of autotrophic organisms on land and in the ocean is of critical importance

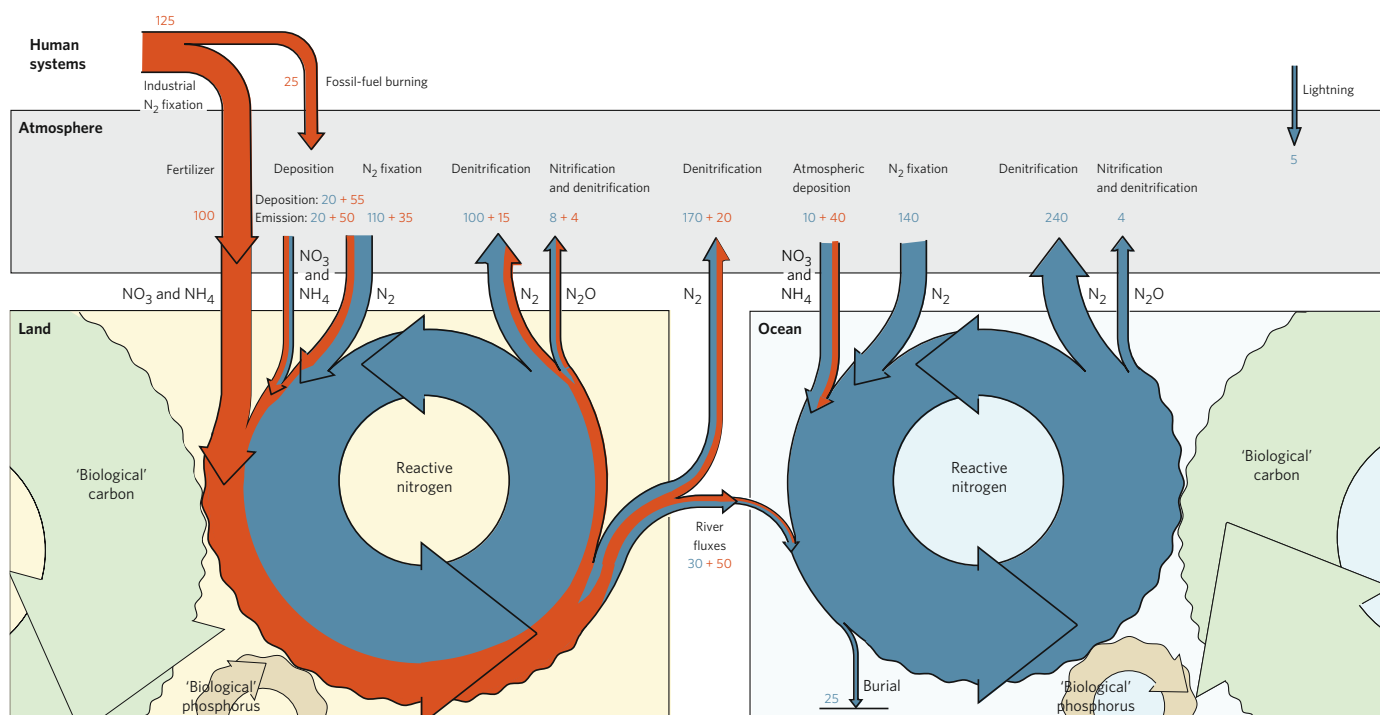
for understanding the global nitrogen and carbon cycles and the Earth system. Given nitrogen's importance in limiting global primary production, about half of which occurs on land and half in the ocean<sup>9</sup>, systematic alterations of the C/N ratios of either marine or terrestrial autotrophs would permit Earth's biosphere to undergo rapid and large changes in productivity without the need to alter the amount of biologically available nitrogen. Such productivity changes would directly affect atmospheric CO<sub>2</sub>, and consequently climate. In contrast, if the C/N ratios of autotrophic organisms were constrained to vary only within narrow bounds, Earth's productivity would be relatively tightly coupled to the amount of biologically available nitrogen, permitting productivity to vary only within restricted limits unless there were processes that altered the amount of biologically available nitrogen.

### Changing reactive nitrogen inventories

Biological nitrogen fixation and denitrification (which refers here to all processes that convert reactive forms of nitrogen to molecular nitrogen (N<sub>2</sub>), which cannot be used directly as a nitrogen source by most organisms) are the most important natural processes that could alter the amount of reactive nitrogen in the Earth system, and hence alter the global carbon cycle and climate, without changing the C/N ratio of autotrophs (Fig. 1).

In the ocean, the magnitude of biological nitrogen fixation and denitrification and the corollary question of how well these two processes balance each other are currently hotly debated. Current estimates of the marine nitrogen budget arrive at either more-or-less balanced budgets (albeit with large uncertainties)<sup>10</sup> or a very large deficit, driven primarily by a much larger denitrification estimate<sup>11</sup>. Observations so far are not adequate to clearly refute either estimate, but there is no doubt that the marine nitrogen cycle is very dynamic, with a residence time for reactive nitrogen — the time for the total pool of reactive nitrogen to be turned over — of less than 3,000 years<sup>10</sup>.

One is immediately tempted to ask what couples biological nitrogen fixation and denitrification in the ocean, so that the amount of fixed nitrogen in the ocean remains relatively stable over timescales longer than a few thousand years. Although many hypotheses have been put



**Figure 1 | Depiction of the global nitrogen cycle on land and in the ocean.** Major processes that transform molecular nitrogen into reactive nitrogen, and back, are shown. Also shown is the tight coupling between the nitrogen cycles on land and in the ocean with those of carbon and

phosphorus. Blue fluxes denote 'natural' (unperturbed) fluxes; orange fluxes denote anthropogenic perturbation. The numbers (in Tg N per year) are values for the 1990s (refs 13, 21). Few of these flux estimates are known to better than  $\pm 20\%$ , and many have uncertainties of  $\pm 50\%$  and larger<sup>13,21</sup>.

forward, current evidence suggests that the marine phosphorus cycle is crucial in stabilizing the marine nitrogen cycle<sup>12</sup>, with other factors such as light, temperature and iron availability having a modulating effect. This hypothesis essentially makes the marine nitrogen cycle a slave to that of phosphate, making phosphate the ultimate limiting nutrient — that is, the nutrient that puts an upper limit on marine productivity and the ocean carbon cycle on timescales of thousands of years and longer.

In contrast with the marine realm, relatively few studies have attempted to scale up local estimates of biological nitrogen fixation and denitrification in terrestrial systems to the global scale<sup>13</sup>, making the terrestrial reactive nitrogen budget as tentative as that of the ocean. When all estimated losses of nitrogen from terrestrial systems are subtracted from estimates of nitrogen inputs to these systems, the balance — which includes the accumulation of reactive nitrogen in the system — is statistically indistinguishable from zero. However, there are such large uncertainties in the individual estimates that estimates of accumulation made by such difference methods are meaningless, other than to say that it is occurring. As a result of a somewhat smaller pool size, the total reactive nitrogen on land is turned over even more rapidly than that in the ocean, having a mean residence time of only about 500 years.

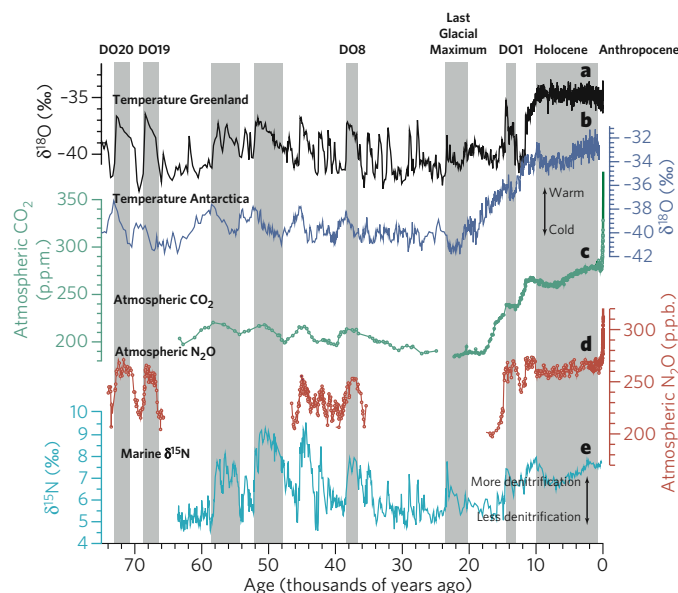
Nearly half of global terrestrial denitrification occurs in freshwater systems<sup>14</sup>, with most of the reactive nitrogen that is denitrified coming from the land. Thus, terrestrial nitrogen cycling is characterized by a strong lateral transport component, which brings reactive nitrogen from the land, where sources of reactive nitrogen tend to exceed local denitrification, into freshwater systems, where the opposite is the case. There, most of the land-derived reactive nitrogen is removed, leaving a comparatively small flux of reactive nitrogen entering the ocean<sup>14</sup>. This one-way conveyor prevents the terrestrial nitrogen cycle from having such a tight bidirectional interaction between biological nitrogen fixation and denitrification as is hypothesized to occur in the ocean. Thus, the question of what controls nitrogen fixation and denitrification in terrestrial systems, and what keeps the terrestrial nitrogen cycle in balance on timescales longer than a few thousand years, is even more perplexing than for marine systems.

### Past changes as a guide?

A good test of our knowledge of the global nitrogen cycle and of its interaction with the carbon cycle and climate is the past. In the past million years, Earth's climate has undergone many large swings, to which the global nitrogen cycle has responded sensitively (Fig. 2).

Perhaps the most informative record of the past activity of the global nitrogen cycle is that of atmospheric nitrous oxide ( $\text{N}_2\text{O}$ ), as its concentration is primarily determined by the magnitudes of nitrification (that is, the oxidation of ammonia to nitrite and, subsequently, to nitrate) and denitrification — two central processes of the global nitrogen cycle. Over the past 60,000 years (Fig. 2),  $\text{N}_2\text{O}$  has undergone large and relatively rapid changes that are synchronized with climate variations, with cold periods generally corresponding to low  $\text{N}_2\text{O}$  concentrations, and vice versa. However, the response of  $\text{N}_2\text{O}$  to these climate changes is not linear but is characterized by hystereses and enhanced responses to prolonged climatic perturbations<sup>15</sup>. As the ocean and the land contribute about equally to natural  $\text{N}_2\text{O}$  emissions, both systems could be responsible for these changes in atmospheric  $\text{N}_2\text{O}$ , but attribution has remained elusive so far. Despite this lack of understanding of the underlying processes forcing these changes, the close correspondence between atmospheric  $\text{CO}_2$  levels, temperature and atmospheric  $\text{N}_2\text{O}$  concentrations demonstrate that the nitrogen cycle is closely coupled to variations in the climate system and in the carbon cycle.

Data from the marine environment underscore this coupling. Measurements of the  $^{15}\text{N}/^{14}\text{N}$  ratio of organic nitrogen from marine sediments in the Arabian Sea (Fig. 2) show rapid variations that are remarkably similar to those of atmospheric  $\text{CO}_2$  and climate. These  $^{15}\text{N}/^{14}\text{N}$  variations largely reflect changes in marine denitrification, with high values characterizing periods with elevated denitrification — that is, high losses of reactive nitrogen from the marine realm — potentially leading to a reduction in the strength of marine productivity. Given the correspondence between high



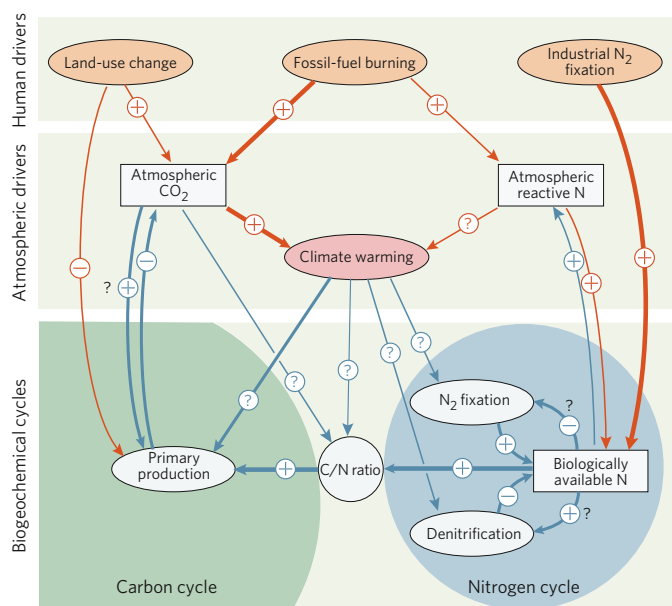
**Figure 2 | Changes in the climate system and the global nitrogen and carbon cycles over the past 75,000 years.** Data are plotted against age before the year 1950, using the Greenland-based GISP2 age scale. **a**, The  $^{18}\text{O}/^{16}\text{O}$  ratio ( $\delta^{18}\text{O}$ ) of ice from Greenland, as a proxy for Greenland temperature<sup>22</sup>. **b**, The  $\delta^{18}\text{O}$  of ice from Antarctica, as a proxy for Antarctic temperature<sup>22</sup>. **c**, Atmospheric  $\text{CO}_2$  concentrations as recorded in air bubbles from various Antarctic ice cores (see ref. 23 for original references) and direct atmospheric measurements since 1958. **d**, Atmospheric  $\text{N}_2\text{O}$  concentrations as recorded in air bubbles from various Antarctic and Greenland ice cores<sup>15</sup> and direct atmospheric measurements since the late 1970s. **e**,  $^{15}\text{N}/^{14}\text{N}$  ratio ( $\delta^{15}\text{N}$ ) of organic nitrogen from a marine sediment core from the Oman margin in the Arabian Sea<sup>16</sup>. ‰, parts per thousand; DO, Dansgaard-Oeschger event; p.p.b., parts per billion; p.p.m., parts per million.

ocean denitrification rates and high atmospheric  $\text{CO}_2$  levels, it has been suggested that changes in the marine nitrogen cycle could be a leading cause of the observed variations in the concentration of atmospheric  $\text{CO}_2$  (ref. 16). Such a nitrogen-based hypothesis to explain the large variations in atmospheric  $\text{CO}_2$  concentrations across the glacial-interglacial periods of the past 650,000 years is tempting, as the causes of these changes are still not clearly identified and represent one of the greatest enigmas of global carbon-cycle research. However, a recent assessment concluded that it is unlikely that changes in the marine nitrogen cycle were the key drivers for the past changes in  $\text{CO}_2$  levels, although they probably contributed to it<sup>10</sup>.

Another key message from the records of the past is that anthropogenic perturbation of the global carbon and nitrogen cycles already pushed these cycles into uncharted territory decades ago, with atmospheric  $\text{CO}_2$  and  $\text{N}_2\text{O}$  now having attained levels that have, almost certainly, not been seen on this planet for the past 650,000 years<sup>2</sup>.

### The future

What will the future hold? Future assessments are rife with uncertainties, but it is difficult to conceive a trajectory of global development up to at least 2050, and possibly beyond, that will not result in increased industrial production of nitrogen-based fertilizers and increased emissions of fossil-fuel  $\text{CO}_2$  (ref. 2). The level that atmospheric  $\text{CO}_2$  will attain in the future depends not only on the rate of anthropogenic emissions, but to a substantial degree on the future behaviour of the Earth system<sup>17</sup>, which so far has helped to mitigate the anthropogenic  $\text{CO}_2$  problem substantially by absorbing roughly half of total  $\text{CO}_2$  emissions<sup>4</sup> (see page 297). With the atmospheric  $\text{CO}_2$  levels currently projected up to 2100, one expects an additional warming of between a few and several degrees Celsius<sup>2</sup>. Thus, there is little doubt that the global nitrogen cycle will come under increasing pressure, not only from direct anthropogenic perturbations but also from the consequences of



**Figure 3 | Nitrogen-carbon-climate interactions.** The main anthropogenic drivers of these interactions during the twenty-first century are shown. Plus signs indicate that the interaction increases the amount of the factor shown; minus signs indicate a decrease; question marks indicate an unknown impact (or, when next to a plus or minus sign, they indicate a high degree of uncertainty). Orange arrows denote the direct anthropogenic impacts, and blue arrows denote natural interactions, many of which could also be anthropogenically modified. Arrow thickness denotes strength of interaction. Only selected interactions are shown.

climate change. At the same time, the response of the global nitrogen cycle to these forcings could have major consequences for the further evolution of climate change. It could have either an enforcing effect, by reducing the ability of the Earth system to absorb anthropogenic  $\text{CO}_2$  (positive feedback), or a reducing effect, by increasing the uptake of anthropogenic  $\text{CO}_2$  (negative feedback).

There are too many possible interactions to assess in this brief article, but some of the interacting drivers of the nitrogen cycle during the twenty-first century are presented in Fig. 3. From the perspective of nitrogen-carbon-climate interactions, the following two processes need special consideration: decoupling of the nitrogen and carbon cycling through changes in the C/N ratios of autotrophs; and changes in the reactive nitrogen inventory of the Earth system through changes in nitrogen fixation (industrial and biological), denitrification or mobilization.

An example for the first process is the recent finding that ocean acidification resulting from the ocean's taking up anthropogenic  $\text{CO}_2$  might lead to an increase in the C/N uptake ratio of marine phytoplankton<sup>18</sup> and enhanced nitrogen fixation<sup>19</sup>. If this tentative result holds up, these changes would make the marine biosphere act as a negative feedback for climate change, as the resulting enhanced fixation of carbon would draw additional carbon from the atmosphere, thus reducing the accumulation of anthropogenic  $\text{CO}_2$  in the atmosphere.

A good example of the second process is the role of the reactive nitrogen inventory in the future productivity of terrestrial ecosystems. The current generation of coupled climate-carbon-cycle models used for making projections of Earth's climate for the remainder of the twenty-first century and beyond<sup>17</sup> do not consider nitrogen limitation of the terrestrial biosphere but generally assume a strong  $\text{CO}_2$  fertilization effect. In several models, the magnitude of this fertilization-induced uptake amounts in the next 100 years to several hundred petagrams of carbon, which requires several thousand teragrams of nitrogen. This amount of reactive nitrogen is clearly not available in the Earth system. Thus, nitrogen limitation is bound to substantially determine the ability of the terrestrial biosphere to act as a  $\text{CO}_2$  sink in the future, although the detailed interactions between increased fertility, C/N ratios in plants and soils and microbial activity are

only poorly understood. The lack of consideration of this whole class of climate-relevant feedbacks in the current Earth system leads to substantial uncertainties in climate-change projections<sup>17</sup>.

Such uncertainties urgently need to be reduced, because major political, societal and economic decisions need to be undertaken if humans are serious in addressing the challenges associated with future climate change. The reduction of these uncertainties requires a major concerted effort that includes the entire set of tools and approaches available to researchers who work in the fields of carbon and nitrogen studies. A particularly pressing need is for ecosystem-manipulation studies that address the interactions of multiple perturbation factors.

Can management of the global nitrogen cycle help to mitigate climate change? Although various options have been proposed in the past, such as fertilization of forests and marine ecosystems, the scientific consensus is that their effectiveness is generally low, and that unintended negative consequences could be serious<sup>20</sup>. Therefore, the best strategy for reducing the potential threat from human activity in the 'Anthropocene' — this modern age in which humans have a significant impact on the Earth system — is to reduce the burning of fossil fuels.

Nicolas Gruber is in the Environmental Physics group, Institute of Biogeochemistry and Pollutant Dynamics, ETH Zurich, Universitätsstrasse 16, 8092 Zurich, Switzerland. James N. Galloway is in the Environmental Sciences Department, University of Virginia, 291 McCormick Road, Charlottesville, Virginia 22904, USA.

- Falkowski, P. G. *et al.* The global carbon cycle: a test of our knowledge of Earth as a system. *Science* **290**, 291–296 (2000).
- Intergovernmental Panel on Climate Change. in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Solomon, S. *et al.*) 1–18 (Cambridge Univ. Press, Cambridge, UK, 2007).
- Galloway, J. N. *et al.* The nitrogen cascade. *Bioscience* **53**, 341–356 (2003).
- Sarmiento, J. L. & Gruber, N. Anthropogenic carbon sinks. *Physics Today* **55**, 30–36 (2002).
- Schimel, D. S. *et al.* Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems. *Nature* **414**, 169–172 (2001).
- Hyvönen, R. *et al.* Impact of long-term nitrogen addition on carbon stocks in trees and soils in northern Europe. *Biogeochemistry*, doi:10.1007/s10533-007-9121-3 (2007).
- Rabalais, N. N. Nitrogen in aquatic environments. *Ambio* **31**, 102–112 (2002).
- Sterner, R. W. & Elser, J. J. *Ecological Stoichiometry: the Biology of Elements from Molecules to the Biosphere* (Princeton Univ. Press, Princeton, 2002).
- Field, C. B., Behrenfeld, M. J., Randerson, J. & Falkowski, P. Primary productivity of the biosphere: an integration of terrestrial and oceanic components. *Science* **281**, 237–240 (1998).
- Gruber, N. in *Carbon Climate Interactions* (eds Oguz, T. & Follows, M.) 97–148 (Kluwer Academic, Dordrecht, 2004).
- Codispoti, L. A. An oceanic fixed nitrogen sink exceeding 400 Tg N a<sup>-1</sup> vs the concept of homeostasis in the fixed-nitrogen inventory. *Biogeosciences* **3**, 1203–1246 (2006).
- Deutsch, C., Sarmiento, J. L., Sigman, D. M., Gruber, N. & Dunne, J. P. Spatial coupling of nitrogen inputs and losses in the ocean. *Nature* **445**, 163–167 (2007).
- Galloway, J. N. *et al.* Nitrogen cycles: past, present, future. *Biogeochemistry* **70**, 153–226 (2004).
- Seitzinger, S. *et al.* Denitrification across landscapes and waterscapes: a synthesis. *Ecol. Appl.* **16**, 2064–2090 (2006).
- Flückiger, J. *et al.* N<sub>2</sub>O and CH<sub>4</sub> variations during the last glacial epoch: insight into global processes. *Global Biogeochem. Cycles* **18**, 1–14 (2004).
- Altabet, M. A., Hignson, M. J. & Murray, D. W. The effect of millennial-scale changes in Arabian Sea denitrification on atmospheric  $\text{CO}_2$ . *Nature* **415**, 159–162 (2002).
- Denman, K. L. *et al.* in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Solomon, S. *et al.*) 499–587 (Cambridge Univ. Press, Cambridge, UK, 2007).
- Riebesell, U. *et al.* Enhanced biological carbon consumption in a high  $\text{CO}_2$  ocean. *Nature* **450**, 545–548 (2007).
- Barcelos e Ramos, J., Biswas, H., Schulz, K. G., LaRoche, J. & Riebesell, U. Effect of rising atmospheric carbon dioxide on the marine nitrogen fixer *Trichodesmium*. *Global Biogeochem. Cycles* **21**, doi:10.1029/2006GB002898 (2007).
- Austin, A. T. *et al.* in *Interactions of the Major Biogeochemical Cycles* (eds Melillo, J. M., Field, C. B. & Moldan, B.) Ch. 3, 15–46 (Island, Washington DC, 2003).
- Gruber, N. in *Nitrogen in the Marine Environment* 2nd edn (eds Capone, D. G., Bronk, D. A., Mulholland, M. R. & Carpenter, E.) Ch. 1 (Academic, San Diego, in the press).
- Blunier, T. & Brook, E. J. Timing of millennial-scale climate change in Antarctica and Greenland during the last glacial period. *Science* **291**, 109–112 (2001).
- Siegenthaler, U. *et al.* Stable carbon cycle-climate relationship during the Late Pleistocene. *Science* **310**, 1313–1317 (2005).

**Acknowledgements** This work was supported by funds from ETH Zurich. We thank J. Flückiger for helping us with the ice-core records.

**Author Information** Reprints and permissions information is available at [npg.nature.com/reprints](http://npg.nature.com/reprints). Correspondence should be addressed to N.G. ([nicolas.gruber@env.ethz.ch](mailto:nicolas.gruber@env.ethz.ch)).

# A steep road to climate stabilization

Pierre Friedlingstein

**The only way to stabilize Earth's climate is to stabilize the concentration of greenhouse gases in the atmosphere, but future changes in the carbon cycle might make this more difficult than has been thought.**

The present dependence on fossil fuels for energy means that as the demand for energy increases, so does the emission of greenhouse gases. The increasing concentration of these gases in the atmosphere has caused most of the warming observed worldwide over the twentieth century. Moreover, the global average surface temperature is projected to rise by as much as 6.4 °C by the end of the twenty-first century if emissions are not curbed<sup>1</sup>. To avoid the potentially dangerous consequences of such climate changes, the concentration of greenhouse gases in the atmosphere must be stabilized at a level that is 'safe' for society and for the environment — a goal that will require a marked reduction in anthropogenic emissions.

Industrialized countries are currently focusing on 'climate mitigation' policies that, when implemented, will result in reduced emission of greenhouse gases. It was recently proposed that by 2020 each of these countries should reduce emissions to 60–75% of the amount that they emitted in 1990; and by 2050, to 25–50% of 1990 levels<sup>2</sup>. However, no such agreement was reached at the last UN Framework Convention on Climate Change Conference of Parties, held in Bali in December 2007. Nevertheless, these proposals, if acted on soon, are good news. But, to

paraphrase Neil Armstrong, that's one giant leap for policy-makers, but one small step for the global environment.

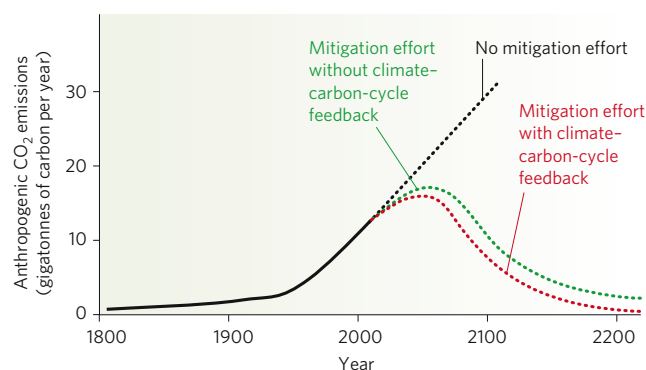
For a start, industrialized countries produce only about 50% of global greenhouse-gas emissions, and the proportion produced by industrializing countries such as China and India is growing. If it is assumed, optimistically, that industrializing countries will not increase their emission rates soon and if industrialized countries follow the above proposal, then global emissions in 2020 will be only 12–20% less than in 1990.

From a glance at the global carbon cycle, it is clear that this reduction will not come close to stabilizing the concentration of greenhouse gases in the atmosphere. At present, deforestation and the combustion of fossil fuels release almost 10 billion tonnes of carbon into the atmosphere each year in the form of CO<sub>2</sub> — the main greenhouse gas. Of this amount, about 4.5 billion tonnes accumulate in the atmosphere, and the rest is absorbed by the ocean and by land-based ecosystems<sup>1</sup>. To stabilize atmospheric CO<sub>2</sub> at the current concentration, emissions would need to be reduced to the amount that is taken up by the ocean and land — about 5.5 billion tonnes, which equates to an immediate 45% reduction in global emissions of CO<sub>2</sub>. This roughly matches the objective proposed

Feedbacks between climate and the carbon cycle mean that emissions of greenhouse gases need to be reduced further than previously thought.



STRINGER SHANGHAI/REUTERS



**Figure 1 | Schematic illustration of past and projected trajectories of anthropogenic CO<sub>2</sub> emissions.** The amount of CO<sub>2</sub> emitted from 1800 to the present is shown in black (solid line). Three projected trajectories of anthropogenic CO<sub>2</sub> emissions are also shown (dashed lines): no effort to reduce emissions (black), and CO<sub>2</sub> stabilization scenarios that do (red) or do not (green) take into account positive feedback between climate and the carbon cycle. It is clear that a greater reduction in emissions will be required to stabilize climate when feedback involving the carbon cycle is considered.

for the industrialized countries for 2050, by which time considerably more CO<sub>2</sub> will have accumulated in the atmosphere.

Moreover, such an immediate reduction would need to be reinforced over time, even if it were achieved. When the concentration of CO<sub>2</sub> in the atmosphere increases, the concentration of the gas in the atmosphere is greater than the concentration in the upper ocean, creating a net flux of CO<sub>2</sub> from the air to the ocean. But, if atmospheric CO<sub>2</sub> concentrations stabilized, the average concentration in the ocean would slowly increase to match the concentration in the atmosphere, so uptake by the ocean would eventually cease. Thus, the immediate 45% reduction in global emissions would no longer be enough to keep CO<sub>2</sub> concentrations constant.

In fact, climate stabilization might be even more complex. Recent observations and simulations indicate that the current uptake of atmospheric CO<sub>2</sub> might be adversely affected by climate change. Careful measurements of the airborne proportion of anthropogenic emissions (that is, the proportion that remains in the atmosphere) show a small increasing trend in the past 50 years<sup>3</sup>. Therefore, the proportion of anthropogenic CO<sub>2</sub> absorbed by the ocean and the land is becoming smaller. The Southern Ocean might be responsible for this reduction, because changes in ocean-surface winds seem to have decreased the amount of CO<sub>2</sub> taken up by surface waters in this region in recent years<sup>4</sup>.

Furthermore, simulations carried out with coupled climate and carbon-cycle models indicate that changes in climate will result in even greater reductions in the ability of land and the ocean to absorb anthropogenic CO<sub>2</sub> by the end of the twenty-first century<sup>5</sup>. These simulations suggest that the combination of warming and drying will limit photosynthesis by plants and stimulate the decomposition of organic matter in soil, reducing the capacity of land-based ecosystems to store carbon (see page 289). In addition, it is widely thought that global warming will result in slower ocean circulation, leading to a decrease in the amount of carbon that is exported from the surface to the deep ocean and thereby reducing the flux of carbon from the air to the ocean. So it seems that future warming will reduce carbon sinks, leaving more CO<sub>2</sub> in the atmosphere and leading, in turn, to greater warming.

This positive-feedback loop has implications for the pathway to stabilizing the concentrations of atmospheric greenhouse gases. If land-based and ocean ecosystems store less carbon than is expected in the future, then a greater effort will be needed, in terms of reducing anthropogenic emissions, to achieve a given concentration of atmospheric CO<sub>2</sub>. The potential importance of this effect is illustrated by simulations carried out for the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC). These simulations indicate that to stabilize atmospheric CO<sub>2</sub> concentrations at 450 parts per million (generally accepted as 'safe') by 2100, cumulative emissions in the twenty-first century need to be reduced by a further 30% when this feedback is taken into account (Fig. 1).

Future policies aimed at stabilizing climate at a safe level will have to take many factors into consideration: the risks and associated financial costs of adapting to climate change; the risks of positive climate and carbon-cycle feedbacks reducing the efficiency of emission-reduction strategies; and the financial costs of reducing emissions. With the aim of informing such policies, the next assessment by the IPCC will explore various scenarios in which emissions are mitigated, including trajectories of emissions over time that result in stabilization of greenhouse-gas concentrations. These scenarios will be used by the climate research community to estimate the extent of future climate change, as well as its impact and the adaptations that might be required. This process differs fundamentally from past assessments by the IPCC, for which climate projections were based on non-mitigated emissions scenarios involving steady increases in greenhouse-gas concentrations over the twenty-first century.

This environmentally concerned view needs to be taken up and followed through by a succession of post-Kyoto regulations in the coming decades that lead to larger and larger reductions in greenhouse-gas emissions and eventually to stabilization of Earth's climate in a state that is safe for society and the environment. There is, unfortunately, no mystery: to stabilize climate, the concentration of greenhouse gases in the atmosphere must be stabilized, and to do so — given the limited capacity of the natural environment to absorb these gases — anthropogenic emissions will eventually need to be reduced to zero.

Pierre Friedlingstein is at the Institute Pierre Simon Laplace, Laboratory of Climate and Environment Sciences, CEA-Saclay, 91191 Gif-sur-Yvette, France.

1. Solomon, S. *et al.* (eds) *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, Cambridge, UK, 2007).
2. The Vienna Climate Change Talks 2007, United Nations Framework Convention on Climate Change ([http://unfccc.int/meetings/intersessional/awg\\_4\\_and\\_dialogue\\_4/items/3999.php](http://unfccc.int/meetings/intersessional/awg_4_and_dialogue_4/items/3999.php)), and the United Nations Framework Convention on Climate Change, COP13, Bali, 2007 ([http://unfccc.int/meetings/cop\\_13/items/4049.php](http://unfccc.int/meetings/cop_13/items/4049.php)).
3. Canadell, J. G. *et al.* Contributions to accelerating atmospheric CO<sub>2</sub> growth from economic activity, carbon intensity, and efficiency of natural sinks. *Proc. Natl Acad. Sci. USA* **104**, 18866–18870 (2007).
4. LeQuéré, C. *et al.* Saturation of the Southern Ocean CO<sub>2</sub> sink due to recent climate change. *Science* **316**, 1735–1738 (2007).
5. Friedlingstein, P. *et al.* Climate-carbon cycle feedback analysis: results from the C4MIP model intercomparison. *J. Clim.* **19**, 3337–3353 (2006).

**Acknowledgements** I thank the Coupled Carbon Cycle Climate Model Intercomparison Project (C4MIP) community for fruitful discussions. The C4MIP project is supported by the International Geosphere Biosphere Program and the World Climate Research Program. This work was supported by the European Community funded project ENSEMBLES.

**Author Information** Reprints and permissions information is available at [npg.nature.com/reprints](http://npg.nature.com/reprints). Correspondence should be addressed to the author ([pierre.friedlingstein@lsce.ipsl.fr](mailto:pierre.friedlingstein@lsce.ipsl.fr)).

# Small-scale cloud processes and climate

Marcia B. Baker & Thomas Peter

**Clouds constitute the largest single source of uncertainty in climate prediction. A better understanding of small-scale cloud processes could shed light on the role of clouds in the climate system.**

Clouds control Earth's weather and regulate its climate<sup>1,2</sup>. They cool Earth's atmosphere by reflecting incoming visible-wavelength solar radiation and warm its surface by trapping outgoing infrared radiation. Clouds produce the rain and snow that dominate Earth's weather and shape Earth's landscapes and vegetation zones.

The large-scale effects of clouds are difficult to characterize accurately because they result from processes that occur on very small scales<sup>3</sup>. Small particles, ranging in size from nanometres to hundreds of micrometres, are strongly affected by updraughts and downdraughts and turbulent mixing on scales of metres to kilometres. Figure 1 shows schematically how large-scale cloud properties depend on small-scale processes. Submicrometre aerosol particles produced by natural processes, such as dust storms, and by anthropogenic processes, such as burning of wood and fuel, constitute the nuclei on which water droplets and ice crystals form in a cloud. Traditionally thought to consist mainly of sulphates<sup>4</sup>, aerosols are now known to have a much more varied composition. Cloud particles in the form of water droplets or ice crystals then grow by taking up water vapour. The radiative property of clouds depends on the size and the number of cloud particles. If the cloud particles reach tens of micrometres, they fall rapidly, colliding with one another to form rain.

The rates at which cloud particles form, grow and fall out of clouds depend on the concentrations, sizes and chemical compositions of the aerosol particles. They also depend on the humidity, temperature and vertical velocity of the air, and on the fluctuations in these parameters (over a distance of 100 metres to several kilometres). Anthropogenic modification of the concentrations and/or chemical compositions of aerosol particles might therefore influence cloud development, weather and climate. An example of this is shown in Fig. 2; in this satellite photograph of clouds over the Atlantic Ocean, the thin white lines crossing the image are bright clouds consisting of small drops that form on the particles emitted by ships, a particularly vivid demonstration of human activity altering the reflectivity of Earth.

## Recent developments

Over the past few decades, the ability to observe small-scale cloud phenomena has improved markedly. Sophisticated laboratory equipment allows the observation of individual micrometre-sized particles suspended in the air. In addition, satellite-borne instruments can now detect, and to some extent identify, cloud and aerosol particles. With these developments, there have been incredible achievements — but new challenges have also been presented.

Remarkable progress has been made in understanding how aerosol particles modify droplet freezing in clouds. Freezing has tremendous climatic effects because it is often the first step in rain formation. It modifies the rate of cloud ascent, and freezing in the upper troposphere creates cirrus clouds, which are thin clouds (composed of ice crystals) that are effective at trapping outgoing radiation.

It might seem surprising that there is more to learn about freezing. Although bulk freezing near 0 °C is well understood, water in tiny droplets can persist as a supercooled liquid at much lower temperatures.

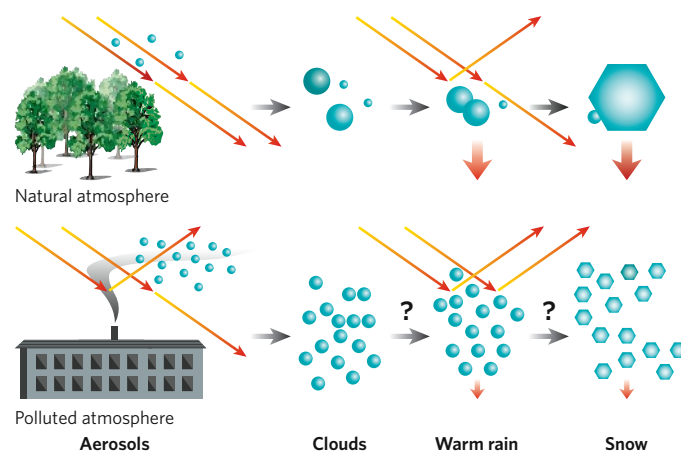
It was traditionally thought that formation of ice in clouds required solid

aerosol particles known as ice nuclei. But it is now known that at temperatures below −35 °C, most ice forms spontaneously through 'homogeneous' freezing in aqueous droplets that contain no foreign particles. The rate of this process depends on, and can be predicted from, air temperature, humidity and small-scale vertical air motions, but is rather insensitive to the chemical composition of the pre-existing aqueous aerosol droplets<sup>5</sup>.

By contrast, ice formation at temperatures between 0 °C and −35 °C can be initiated only through heterogeneous nucleation. Ice nuclei such as mineral-dust particles were originally thought to be inert. But laboratory studies have shown that physicochemical transformation of other aerosol particles, such as those containing organic material, can modify how efficiently ice nuclei function as freezing substrates. This, in turn, makes it difficult to identify the origin of the ice nuclei involved in cloud formation and to predict the climatic role of clouds.

## Current challenges

Because cloud processes are sensitive to the concentration and type of aerosol particle, a major and controversial focus of recent



**Figure 1 | Interactions of aerosol particles with clouds and the consequences for cloud development.** In the natural (non-polluted) atmosphere, the concentrations of aerosol particles are generally low, and the clouds that form on these particles have relatively low droplet and/or ice-crystal concentrations. The increased concentrations of aerosol particles present in polluted atmospheres might lead to the formation of clouds with high concentrations of droplets or ice crystals. Such clouds are expected to reflect more radiation than clouds with fewer droplets<sup>6</sup>, as is corroborated by an increasing body of observational evidence. High cloud-particle concentrations can lead to smaller droplet or crystal sizes and therefore to reduced particle fall speeds (shown by the vertical red arrows; the larger the arrow, the faster the fall). This effect, acting alone, would increase the rate of fallout of precipitation (snow or rain) in non-polluted environments. By contrast, in polluted environments, it would tend to increase the time for which the cloud particles remain suspended<sup>7</sup>. However, observational evidence for aerosol effects on large-scale changes in cloud lifetimes and precipitation patterns is still lacking (as shown by the question marks).



**Figure 2 | Satellite photograph of low clouds over the Atlantic Ocean.** The thin white lines are locally enhanced clouds formed in tracks marking the effluent from smokestacks on passing ships. (Image courtesy of J. Desclotres, MODIS Rapid Response Team, NASA/GSFC, Greenbelt, Maryland.)

cloud research has been the attempt to quantify the extent to which anthropogenic aerosol particles are modifying cloud properties on a global scale. As indicated in Figs 1 and 2, high concentrations of aerosols can increase the brightness of clouds and their ability to reflect solar radiation, a process that is moderately well understood<sup>6</sup> but not well quantified globally. Moreover, it has been suggested that an increased aerosol concentration alters large-scale patterns in cloud lifetimes and precipitation<sup>7</sup>, but these effects of aerosol particles are highly uncertain at present<sup>1</sup>.

New observational programmes<sup>1</sup> and numerical model approaches<sup>8</sup> will be required to pin down the effect of anthropogenic aerosol particles on large-scale cloud properties. Such programmes should include dedicated laboratory investigations of cloud-particle formation and field studies that measure small-scale parameters and follow cloud development over extended periods of time. These process-oriented approaches need to be tightly linked with satellite and surface-based networks that monitor the important cloud variables with sufficient precision and stability to quantify accurately any changes over the coming years to decades.

These improvements will provide exciting intellectual and practical benefits as scientists become increasingly able to predict the development of individual clouds and cloud systems, which has been the goal of much atmospheric research over the past 50 years. Future research into small-scale cloud processes will yield new insights into the large-scale phenomena that characterize Earth's climate. ■

Marcia B. Baker is in the Department of Earth and Space Sciences, University of Washington, Seattle, Washington 98195, USA. Thomas Peter is at the Institute for Atmospheric and Climate Science, ETH Zürich, 8092 Zürich, Switzerland.

1. Solomon, S. *et al.* (eds) *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, Cambridge, UK, 2007).
2. Collins, W., Colman, R., Haywood, J., Manning, M. R. & Mote, P. The physical science behind climate change. *Sci. Am.* **297**, 64–73 (2007).
3. Baker, M. B. Cloud microphysics and climate. *Science* **276**, 1072–1078 (1997).
4. Charlson, R. J. & Wigley, T. M. L. Sulfate aerosol and climatic change. *Sci. Am.* **270**, 48–57 (2004).
5. Koop, T., Luo, B. P., Tsias, A. & Peter, T. Water activity as the determinant for homogeneous ice nucleation in aqueous solutions. *Nature* **406**, 611–614 (2000).
6. Twomey, S. Influence of pollution on the short-wave albedo of clouds. *J. Atmos. Sci.* **34**, 1149–1152 (1977).
7. Albrecht, B. A. Aerosols, cloud microphysics and fractional cloudiness. *Science* **245**, 1227–1230 (1989).
8. Lohmann, U., Quaas, J., Kinne, S. & Feichter, J. Different approaches for constraining global climate models of the anthropogenic indirect aerosol effect. *Bull. Am. Met. Soc.* **88**, 243–249 (2007).

**Acknowledgements** M.B.B. is grateful to R. Wood and G. Raga for helpful comments. T.P. thanks the European Commission and the Swiss National Foundation for financial support.

**Author Information** Reprints and permissions information is available at [npg.nature.com/reprints](http://npg.nature.com/reprints). Correspondence should be addressed to M.B.B. ([marcia@ess.washington.edu](mailto:marcia@ess.washington.edu)).

# Earth science and society

Frank Press

**The unique set of challenges that face humankind today mean that it is more essential than ever that Earth scientists apply their understanding of the planet to benefit society and that society invite them to do so.**

In a single sentence of a speech to the Royal Society, in London, in 1988, Margaret Thatcher succinctly connected science to the creation of social wealth when she said: “the value of Faraday’s work today must be higher than the capitalization of all the shares on the stock exchange.” Add a few other examples of the work of scientists that has transformed society, such as the Green Revolution in world agriculture, the transistor revolution that opened closed societies to change and the biomedical revolution set off by molecular biology, and the benefits of science to society take on real meaning. The Earth sciences have a unique role in this regard, which was underscored by the twentieth-century US historian Will Durant when he is said to have cautioned: “Civilization exists by geological consent, subject to change without notice.” Today, Durant might add a few new vulnerabilities faced by civilization, which I comment on later in this article.

Engagement by scientists in societal matters is not without its problems. An essential element of a democratic society is the accountability of elected officials, who make the final decisions but are answerable to the people. When a serious social problem is addressed, however — one that involves technical matters — scientists are frequently called on for advice. More often than not, the available data are incomplete, the issue is politically charged and the scientists must hold to the integrity of the scientific process in the face of their own personal biases and possible conflicts of interest. In such circumstances, scientists can help decision-makers by describing a range of possible outcomes — assuming there is enough information to do so. However, there are crisis situations where default judgments are needed: that is, where decisions must be made and there is not enough time for more years of research. In this case, I for one would prefer to solicit the views of the most qualified experts in the field in full knowledge of the possible difficulties noted above. Climate change may be such an issue. Crispin Tickell, a former British ambassador to the United Nations, argued the issue this way<sup>1</sup>: “Scientists should be much braver ... I think this ethics argument — should they speak or shouldn’t they — is a lot of nonsense. Scientists cannot promise certainty any more than economists can when they call for changes in taxes or interest rates. Uncertainty is part of the human condition. Caution, in any case, may in reality be recklessness. We must always look at the cost of doing nothing.”

In negotiating the conditions for my position as science adviser to US president Jimmy Carter in early 1977, I learned that he selected me because I was an Earth scientist. To me this signalled his estimate of the important issues he would have to confront in his term of office.

I asked the president for the authority to convene panels of experts to sort through technical issues that might relate to a presidential decision, and this proved to be an important mechanism for providing the counsel of the ranking specialists in a field. On occasion, the president, who was technically competent, chose not to follow our advice, but he respected the process and, where appropriate, he explained the political rationale for his decision.

There are numerous examples of both the contributions that Earth scientists have made to society, and of the effects that society has had on the disciplines within Earth science, and I discuss just a few of them here.

## Natural resources

Just about everything we use — our metals, many of our chemicals, our building materials, silicon for our transistors, our energy resources — comes from the ground. These resources are discovered by geologists who tell us how they were formed, how to find them, and that they will not last forever. Unfortunately, mining can be a dirty business that ravages the environment. Environmental scientists now counsel conservation, recycling and substitution as alternatives to the mining of diminishing resources.

Fresh water, a life-sustaining resource, is faced with growing chemical pollution together with increasing demand. A new threat reported by glaciologists is the retreat of glaciers with the beginnings of global warming. They alert us that this will reduce Earth’s water-storage capacity and seasonal freshwater run-off in many regions, threatening water supplies for drinking and irrigation.

## Living on a violent planet

The same forces that have made our planet so uniquely conducive to life by providing us with continents, oceans and a beneficent atmosphere have also made it a violent planet subject to earthquakes, tsunamis, volcanic eruptions, landslides and floods. Earth scientists share with public authorities the responsibility for showing humankind how to live with these natural hardships and minimize the loss of life and property. In all these cases, they do this by public education, recommending intelligent land use together with regulations that require disaster-resistant design of buildings and other structures. In the case of natural disasters, science can provide early warning with increasing reliability. Earthquake prediction still remains an elusive goal, but real-time seismology is offering hope of improved mitigation (see page 271).



## Climate change

Worrisome as these natural threats are, humankind itself has now become a new troubling force that competes with geology in its power to change our planet. Our ability to alter the chemistry of the atmosphere and thereby change global climate now compares with the natural swings in climate found in the geological record extending back in time over millions of years (see page 279). This is an awesome responsibility because of the profound consequences for humankind and all other living species. In 1896, Swedish scientist Svante Arrhenius calculated that doubling the carbon dioxide ( $\text{CO}_2$ ) content in the atmosphere would raise Earth's temperature by 5–6°C. He proposed that the release of  $\text{CO}_2$  by the combustion of coal would produce global warming. At long last, more than 100 years after Arrhenius's warning, Earth scientists have finally won over most of the world's political leaders to the view that the increased emission of greenhouse gases caused by human activity is responsible for measurable levels of global temperature rise since the mid-twentieth century. The scientists were able to present evidence of troublesome changes in physical and biological systems that could be observed. They could cite detailed observations of receding glaciers, reduced sea-ice cover on the Arctic Ocean, more frequent extreme weather events, early-blooming trees and acidifying oceans. This cause-and-effect linkage was the stunning message in a report by a scientific panel appointed by the United Nations. Hundreds of climate experts from 120 governments contributed to this statement, issued in 2007 by the UN Intergovernmental Panel on Climate Change (IPCC)<sup>2</sup>, and have been rewarded for their efforts by a share in the 2007 Nobel Peace Prize — arguably the highest recognition that scientists can receive for a contribution to society.

In addition, many leading climate scientists are taking what is for them an unusual but necessary action. They are 'going public': that is, expressing in public forums their anxiety about the possible disastrous consequences by the end of this century of unchecked global warming. They are rousing the general public, making this an economic and ethical issue for many business leaders and a political issue for the governments of many countries. Climate experts are now being joined by the many political leaders they have briefed in arguing that even in the absence of absolute certitude (which does not exist for any scientific theory), reduction in greenhouse-gas emissions is mandated because of the non-trivial possibility that global warming could trigger disastrous social and environmental changes. Climate change is an example of a problem faced by scientist-advisers in counselling governments when the issue is politically charged and the early data are incomplete. These scientists persisted, however; the flow of observations and computations

buttressed their case, and they are now forcing economic and political action.

## The ozone hole

Perhaps the most successful example of advice by Earth scientists informing government policy is that of the Montreal Protocol, an international agreement that became effective in 1989 to control the production of industrial chemicals that threatened to destroy the ozone layer in the stratosphere. The rapidity of negotiation and implementation following the publication of the scientific data was remarkable. In 1995, atmospheric chemists Paul Crutzen, Sherwood Rowland and Mario Molina were awarded the Nobel Prize in Chemistry for their work more than two decades earlier on the formation and decomposition of ozone. Molina and Rowland<sup>3</sup> had proposed that a class of normally harmless, commonly used industrial compounds called chlorofluorocarbons, or CFCs, could drift up to the stratosphere. There, photodissociation of the CFCs in a catalytic reaction could produce atomic chlorine that would destroy ozone. Earth's ozone layer, the protective shield that filters cell-damaging solar ultraviolet radiation from reaching the biosphere, could be thinned. In the 1980s, when Earth scientists and others were trying to gain public attention for a possible environmental disaster, a highly placed US government official offered advice that ranks with Marie Antoinette's counsel to starving Parisians: "Let them eat cake." He proposed that as a cheap and effective solution people should wear hats and sunglasses and use sunscreen. Fortunately, ozone depletion over Antarctica was discovered by the British Antarctic Survey in 1985. In the following year, a team of international scientists led by Susan Solomon of the National Oceanic and Atmospheric Administration made *in situ* measurements in the 'ozone hole'. The chemistry of the ozone hole was confirmed. With this evidence, wiser political voices prevailed, and a treaty was rapidly negotiated. By 2007, some 191 countries had ratified the Montreal Protocol, which now envisages the complete phasing out of ozone-depleting substances.

## The nuclear test-ban treaty

Nuclear weapons cannot be developed with confidence that they work without testing. An enforceable ban on testing would thus be a powerful deterrent both to the proliferation of states with nuclear arsenals and to concealed advances in weapons development by nuclear-capable states. We would not be as close as we are today to such a ban without either the work of Earth scientists or the influence of society on the field of seismology. For some 40 years, the United States, Russia and other countries with nuclear weapons have been trying to reach agreement on methods to verify compliance with a test-ban treaty by developing a reliable tool to detect clandestine underground testing of nuclear weapons. Seismic detection of explosions was the obvious technology,

Melting glaciers: one of the factors leading Earth scientists to rouse governments and the general public to take action against climate change.



Rajendra Pachauri accepts the 2007 Nobel Peace Prize on behalf of the IPCC.



Earth scientists contributed to early negotiations for a nuclear test-ban treaty.

but in the early years of negotiations over a treaty, the field of seismology was insufficiently developed to do the complete job of detecting a nuclear explosion, locating it and stating with confidence that the event was an explosion and not an earthquake. That was the driver that transformed the tiny academic research field of seismology into a military-industrial-academic complex that would expose seismologists to the seductions of huge funding increases, co-option by government officials with political agendas, distortion of their research priorities and biased selection of data in publications and testimony.

The first negotiations for a test-ban treaty consisted of several meetings of US, British and Soviet scientists in Geneva, beginning in 1958 (in what follows, I draw on the excellent descriptions of the early history of the nuclear test-ban negotiations in refs 4 and 5). The government of the Soviet Union was leery of foreign inspectors roaming freely in their country in search of evidence for clandestine tests, and Soviet seismologists presented seismological data that supported this political policy, claiming that their seismic networks could easily detect even small detonations of chemical explosives at distances of hundreds of kilometres. The US government thought the Soviets capable of and willing to evade a treaty, and US government scientists presented apparently contradictory evidence of how difficult it was to observe seismic waves generated by the much larger underground nuclear explosions in the United States. They maintained that they would need many seismic stations in the Soviet Union, as well as inspections, to monitor clandestine underground nuclear explosions. Each side suspected the motives of the other's scientists in presenting seemingly slanted evidence in support of their government's political position, but subsequent scientific work revealed that differing regional geology could account for the contradictions. It turned out that the ancient, colder (having a lower geothermal gradient) crustal and mantle rocks of the Eurasian continental shield are more effective at generating and propagating seismic waves from an explosion than are the rocks under the Nevada test site, which sits in a geologically younger region where conditions tend to muffle seismic waves.

To avoid disruption of the negotiations because of the conflicting technical positions of the delegations, the administration of President Eisenhower launched a research programme in seismology called Vela Uniform. An advisory panel was appointed by the president's science adviser James Killian of the Massachusetts Institute of Technology to prepare a research plan. The panel, chaired by science administrator Lloyd Berkner, consisted of 14 members, of whom 9 were distinguished university professors, including 6 of the nation's leading academic Earth scientists. The highly respected Advanced Research Projects Agency (ARPA) of the Department of Defense was designated to manage Vela Uniform, and the Berkner Report set its research agenda. Kai-Henrik Barth reported<sup>4</sup>: "Vela Uniform supported almost every US seismologist and even a number of foreign scientists during the 1960s. From 1959 to 1961, funding for seismology increased by a factor of 30 and remained at this level for the better part of the 1960s." Of great importance to the development of seismology is the fact that the government managers of



Discovery of the ozone hole at Halley Research Station led to a ban on CFCs.

these research funds knew how to find and support the best scientists and provided them wide latitude in the selection of their own research topics, knowing that this was in the best long-term interest of the government.

As far as seismology is concerned, fears about militarization of the field during the cold war, and distortion of the research agenda into narrow sectors of special interest to government patrons, never materialized. On the contrary, the US government's generous support of academic Earth scientists with few limitations over the decades not only led to the development of many advanced methods for differentiating between nuclear explosions and earthquakes but also enabled seismologists to make extraordinary contributions to the study of plate tectonics and to the unravelling of the dynamics of Earth's internal heat engine.

The seismological methods that were developed also had a crucial role in facilitating the adoption of the Comprehensive Nuclear Test-Ban Treaty by the United Nations in 1996, by giving states confidence that compliance with the treaty could be verified. A global international monitoring system of 170 seismic stations, which should be capable of detecting and identifying nuclear explosions as small as 1–2 kilotonnes (ref. 6), is now being installed as part of the treaty. This should be sufficient to inhibit any rogue nation from secretly developing a nuclear weapon. I am sure that it will also lead to new discoveries about Earth's interior and provide useful data for early warning of earthquakes, tsunamis and volcanic eruptions in remote regions.

Earth scientists should be proud of the contributions to society they are making in the course of applying and advancing their science. The wider application of old knowledge still serves many purposes, including lessening the destruction of natural disasters. The latest challenge is to apply the new understanding of our planet that has been uncovered by research to halt and reverse the environmental damage inflicted by humankind.

Frank Press is president emeritus of the US National Academy of Sciences and institute professor emeritus at the Massachusetts Institute of Technology. He is currently a director of the Washington Advisory Group of the Law and Economics Consulting Group.

1. Pearce, F. The green diplomat. *New Sci.* no. 1813 38 (1992).
2. Solomon, S. et al. (eds) *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, Cambridge, UK, 2007).
3. Molina, M. J. & Rowland, F. S. Stratospheric sink for chlorofluoromethanes: chlorine atom catalysed destruction of ozone. *Nature* **249**, 810–812 (1974).
4. Barth, K. The politics of seismology: nuclear testing, arms control, and the transformation of a discipline. *Soc. Stud. Sci.* **33**, 743–781 (2003).
5. Richards, P. G. & Zavales, J. in *Monitoring a Comprehensive Test Ban Treaty* (eds Husebye, E. S. & Dainty, A. M.) 53–81 (Kluwer Academic, Dordrecht, 1996); revised <<http://www.ideo.columbia.edu/~richards/earlyCTBTstory.html>>.
6. Committee on Technical Issues Related to Ratification of the Comprehensive Nuclear Test Ban Treaty, and Committee on International Security and Arms Control, National Academy of Sciences. *Technical Issues Related to the Comprehensive Nuclear Test Ban Treaty* (National Academy of Sciences, Washington DC, 2002).

**Author Information** Reprints and permissions information is available at [npg.nature.com/reprints](http://npg.nature.com/reprints). Correspondence should be addressed to the author (fpress@theadvisorygroup.com).

# Precise auditory–vocal mirroring in neurons for learned vocal communication

J. F. Prather<sup>1</sup>, S. Peters<sup>2</sup>, S. Nowicki<sup>1,2</sup> & R. Mooney<sup>1</sup>

**Brain mechanisms for communication must establish a correspondence between sensory and motor codes used to represent the signal. One idea is that this correspondence is established at the level of single neurons that are active when the individual performs a particular gesture or observes a similar gesture performed by another individual. Although neurons that display a precise auditory–vocal correspondence could facilitate vocal communication, they have yet to be identified. Here we report that a certain class of neurons in the swamp sparrow forebrain displays a precise auditory–vocal correspondence. We show that these neurons respond in a temporally precise fashion to auditory presentation of certain note sequences in this songbird’s repertoire and to similar note sequences in other birds’ songs. These neurons display nearly identical patterns of activity when the bird sings the same sequence, and disrupting auditory feedback does not alter this singing-related activity, indicating it is motor in nature. Furthermore, these neurons innervate striatal structures important for song learning, raising the possibility that singing-related activity in these cells is compared to auditory feedback to guide vocal learning.**

To enable learned vocal communication, the brain must establish a correspondence between auditory and motor representations of the vocalization and use auditory information to modify vocal performance. Individual neurons that display a precise auditory–vocal correspondence could enable auditory activity to be evaluated in the context of the animal’s vocal repertoire, facilitating perception. These neurons could also play an important role in vocal learning, because their motor-related activity could be compared with auditory feedback to modify vocalizations adaptively. Despite their potential importance to learned forms of vocal communication, including human speech, single neurons displaying a precise auditory–vocal correspondence have not been identified.

One major difficulty in identifying auditory–vocal attributes of individual neurons has been the challenge of recording from individual neurons in freely vocalizing animals. Another challenge in characterizing sensory and motor properties of neurons for learned vocalizations, such as human speech, is the dearth of suitable animal models. We overcame these challenges by using a lightweight chronic recording device<sup>1</sup> to sample neural activity in male swamp sparrows (*Melospiza georgiana*), a wild songbird that resembles humans in its dependence on auditory experience to learn its vocal communication signals<sup>2–4</sup>. Individual swamp sparrows sing only a few song types (range: 2–5 song types), each comprising a single trilled, multi-note syllable<sup>5</sup> (Supplementary Fig. 1a), simplifying exploration of the auditory and motor representations of the animal’s vocal repertoire.

We focused our search in the telencephalic nucleus HVC, a structure necessary for singing<sup>6</sup> and normal song perception<sup>7</sup> and where high-level motor and auditory representations of birdsong have been detected<sup>8–12</sup>. HVC contains two distinct populations of projection neurons<sup>13</sup>, including one (HVC<sub>RA</sub>) that innervates song premotor neurons in the robust nucleus of the arcopallium (RA)<sup>6</sup> and another (HVC<sub>X</sub>) that innervates a striatal region of the avian basal ganglia<sup>14</sup> (area X<sup>6</sup>) important to song learning and perception<sup>15,16</sup> (Supplementary Fig. 1b). Multiunit recordings from the HVC of awake songbirds have detected song-related auditory and motor activity<sup>17</sup>, but whether single neurons display both types of activity remains

unknown. Furthermore, single neurons downstream of HVC, in the song premotor nucleus RA, exhibit similar patterns of singing-related and auditory activity, but auditory activity was evident only when the bird was asleep<sup>18</sup>, making it difficult to reconcile this auditory activity with a possible role in communication. To test whether individual HVC neurons display similar patterns of auditory and singing-related activity, we recorded from identified projection neurons in the HVC of awake and freely behaving adult male swamp sparrows during auditory presentation of birdsong and during singing (Supplementary Fig. 1b).

## Auditory properties of identified HVC neurons

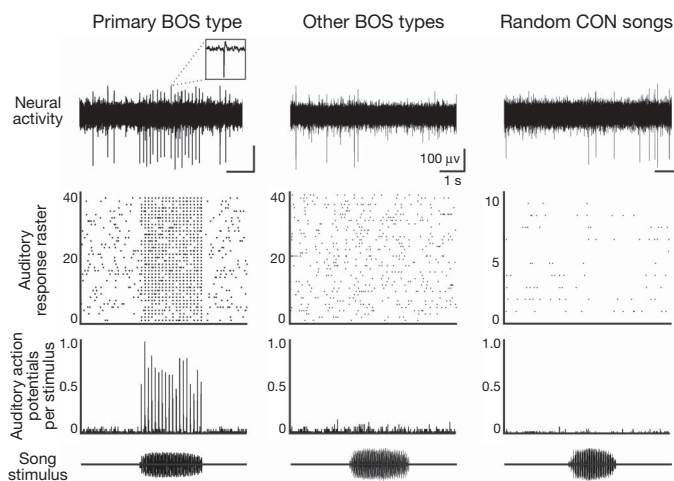
We probed auditory responses of identified HVC neurons by playing a variety of song stimuli, including the bird’s own song types and songs of other swamp sparrows, through a loudspeaker near the bird’s perch. A substantial subset of HVC<sub>X</sub> neurons (21 of 60 HVC<sub>X</sub> cells, 7 birds) responded robustly to song playback (Fig. 1), whereas HVC<sub>RA</sub> neurons were entirely unresponsive (16 HVC<sub>RA</sub> cells, 5 birds; Supplementary Fig. 1c). In a substantial proportion of responsive HVC<sub>X</sub> neurons (16 of 21 cells), auditory activity was selectively evoked by acoustic presentation of only one song type in the bird’s repertoire, defined as the ‘primary song type’, and not by other swamp sparrow songs chosen at random (Fig. 1, Supplementary Fig. 1d, e). The primary song type varied among cells from the same bird, as expected given that each bird produces several song types. Because swamp sparrow song types consist of one syllable trilled many times (Supplementary Fig. 1a), action potential activity evoked throughout song presentation could be plotted as a response to many presentations of a single syllable (see below). This arrangement revealed that action potential activity in HVC<sub>X</sub> neurons occurred at a precise phase relative to syllable onset (s.d. of action potential latency:  $18.34 \pm 14.33$  ms, or  $15.05 \pm 10.85\%$  of syllable duration,  $N = 21$  cells) and was both temporally sparse (action potentials per syllable in which a response occurred:  $1.55 \pm 0.49$ ; action potential burst rate:  $133 \pm 63$  Hz,  $N = 21$  cells) and reliable (probability of activity per syllable:  $0.64 \pm 0.18$ ,  $N = 21$  cells). Thus,

<sup>1</sup>Department of Neurobiology, Duke University Medical Center, <sup>2</sup>Department of Biology, Duke University, Durham, North Carolina 27710, USA.

HVC<sub>X</sub> neurons display auditory responses highly selective in the stimulus domain, typically being activated by only one song type in the bird's repertoire. These auditory responses also are sparse in the time domain, occurring at a precise phase in the syllable of the effective song type.

### Individual neurons are active during listening and singing

To investigate whether auditory HVC<sub>X</sub> neurons also were active during singing, we relied on the tendency of swamp sparrows to countersinging (5 birds, 555 cases of countersong)—this is a territorial singing behaviour triggered by presentation of either the bird's own songs or those of other swamp sparrows (Fig. 2a–c). We exploited this antiphonal behaviour to rapidly assess the auditory and singing-related activity of single neurons in the context of communication. Individual HVC<sub>X</sub> neurons could be active during both listening and singing (Fig. 2a–c;  $N = 7$  cells, 3 birds). Moreover, the most robust singing-related activity in each HVC<sub>X</sub> cell occurred in association with the primary song type, as defined using auditory stimulus presentation (Supplementary Fig. 1e). Notably, the mean timing of singing-related activity, plotted relative to syllable onset, was the same as the mean timing of activity evoked by presentation of the same song type when the bird was not singing (Fig. 3). An auditory–vocal correspondence of this sort was observed in every HVC<sub>X</sub> neuron for which we were able to record activity during both singing and song playback. Further parallels between singing and auditory activity of HVC<sub>X</sub> cells were that action potential activity recruited during singing was reliable (probability of activity per syllable:  $0.91 \pm 0.08$ ,  $N = 7$  cells), persistent throughout the entire song (for example, Fig. 2a, c), and restricted to a limited phase relative to syllable onset (Fig. 3a–c; s.d. of action potential latency:  $7.21 \pm 3.02$  ms,  $N = 7$  cells). One difference in HVC<sub>X</sub> activity between the singing and listening states was that the singing-related activity involved short bursts of action potentials, whereas auditory-evoked activity typically consisted of single action potentials (Fig. 3a, b; action potentials per syllable:  $1.27 \pm 0.21$  auditory,  $2.89 \pm 0.67$  singing,  $P = 0.004$ ; action potential burst rate:  $148 \pm 64$  Hz auditory,  $278 \pm 80$  Hz singing,  $P = 0.01$ ; paired  $t$ -tests,  $N = 7$  cells, 3 birds). In summary, HVC<sub>X</sub> neurons display highly similar, temporally precise



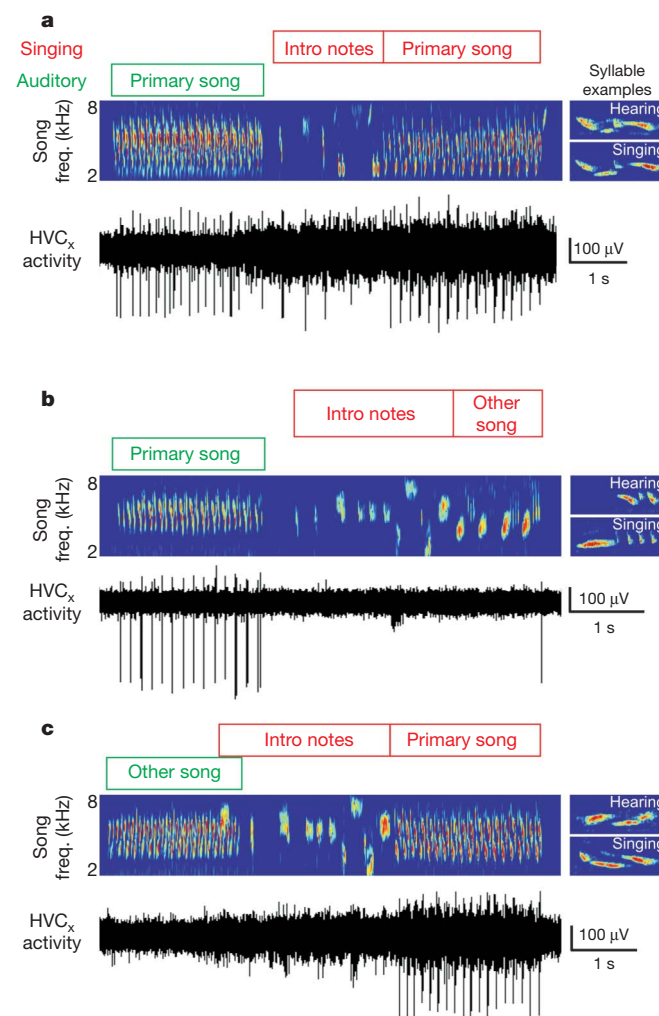
**Figure 1 | In freely behaving swamp sparrows, HVC<sub>X</sub> neurons respond selectively to one song type in the bird's repertoire.** A single song type in the repertoire of the bird's own songs (BOS) typically evoked an auditory response (left, the 'primary BOS type'), whereas other BOS types (middle) and randomly selected songs of conspecific birds (CON, right) were ineffective stimuli (top row, raw data recorded from an HVC<sub>X</sub> neuron during a single stimulus presentation; second row, response raster for multiple presentations; third row, peri-stimulus time histogram (PSTH), 10 ms bin size; bottom row: stimulus oscillogram). Audio files available as Supplementary Information.

patterns of activity while hearing and singing the primary song type, suggestive of a precise sensorimotor correspondence (Fig. 3d).

### Singing-related activity is a corollary discharge

In a simple model of sensorimotor correspondence, motor-related activity should occur before sensory feedback elicited by the action. In this context, the similar action potential timing we observed in HVC<sub>X</sub> cells during singing and listening raises the possibility that the activity during singing was due to auditory feedback. Alternatively, singing-related activity may constitute a corollary discharge of the song motor activity, perhaps providing a motor estimation of auditory feedback<sup>19</sup>. Three observations indicated that HVC<sub>X</sub> activity during singing was motor-related corollary discharge.

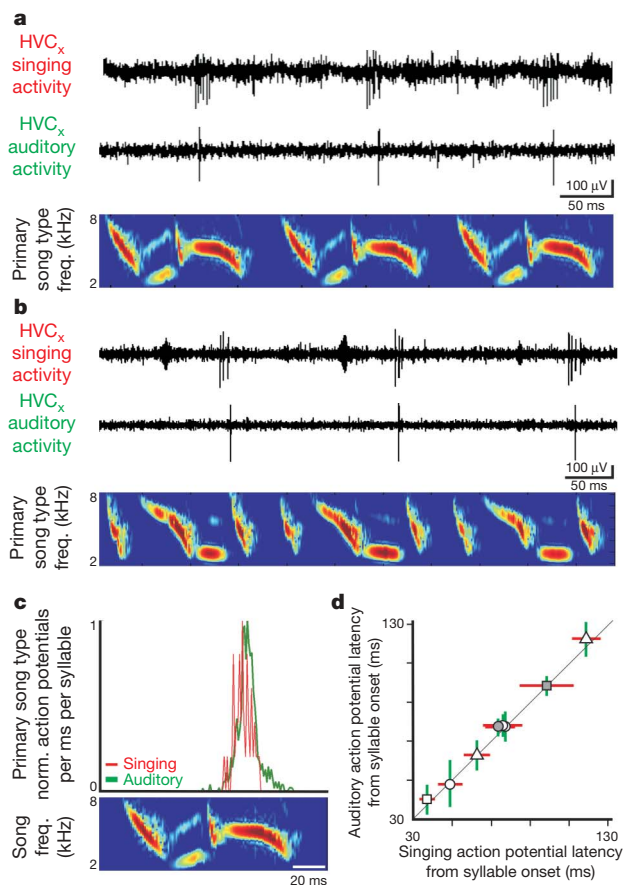
First, we noted that background multiunit activity could increase before singing of any song type (for example, Fig. 2a, c) and, during this 'warm-up' period, the isolated HVC<sub>X</sub> cell's auditory responses to the primary song type were suppressed (Fig. 4a, Supplementary Fig. 2a;  $N = 25$  occurrences, 5 cells, 2 birds). This suppression of auditory activity suggests that HVC<sub>X</sub> neurons switch from an auditory state to an auditory-insensitive motor state several hundred milliseconds



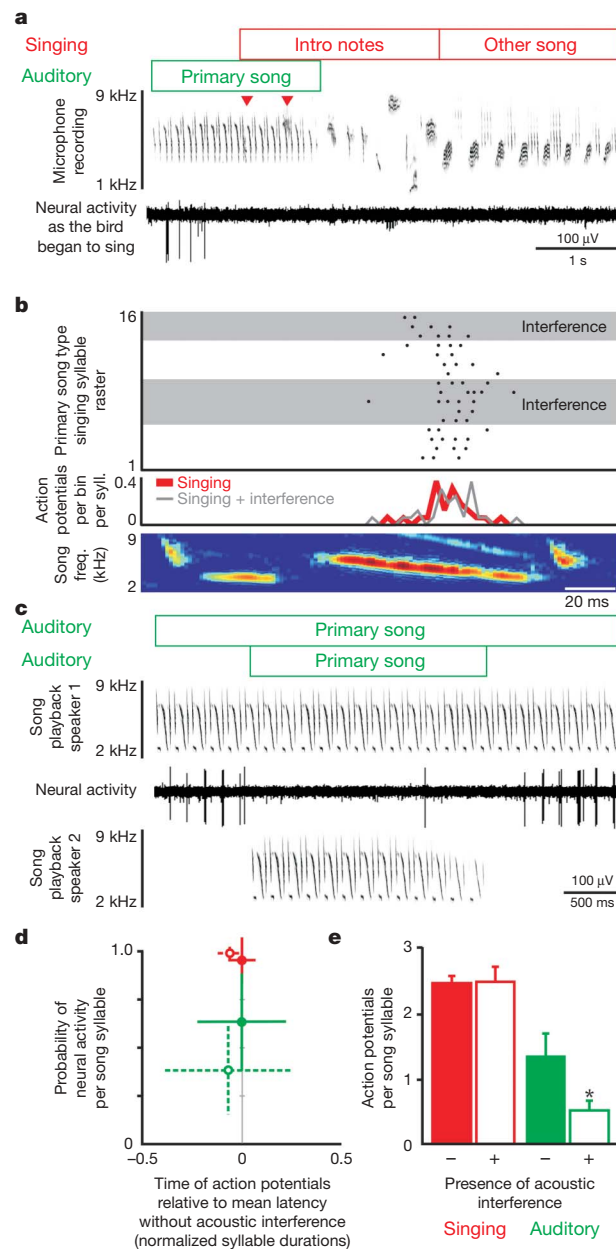
**Figure 2 | Countersinging in response to song presentation reveals auditory and singing-related activity of HVC<sub>X</sub> neurons in the context of communication.** a, In 'matched countersinging'<sup>41</sup>, an HVC<sub>X</sub> neuron was active when the bird heard (green box, left) or sang (red box, right) the primary song type. b, c, In 'unmatched countersinging', another HVC<sub>X</sub> neuron was active when the bird heard (b) or sang (c) the primary song type but was silent as the bird sang (b) or heard (c) other song types. (In a, b, c: top, spectrogram of the acoustic signal; bottom, corresponding electrophysiological recording.) Audio files available as Supplementary Information.

before the onset of singing ( $601 \pm 288$  ms,  $N = 5$  cells, 2 birds). Furthermore, in cases where playback of the primary song type began immediately after the bird stopped singing, auditory-evoked activity remained briefly suppressed ( $\sim 250$  ms, much shorter than reported for HVC multi-unit activity in other species<sup>17</sup>, Supplementary Fig. 2b). Second, there was often a 'secondary song type' for which singing-related activity in an HVC<sub>X</sub> neuron was evident, even though playback of that song type evoked no response from that cell (Supplementary Fig. 2c). Third, in several instances singing of either the primary or secondary song type overlapped playback (N = 4 cells, 3 birds), distorting auditory feedback. However, the singing-related activity pattern was unaffected by such distortion (Fig. 4b, d, e, Supplementary Fig. 2d, e; probability of neural activity:  $P = 0.59$ ; mean latency:  $P = 0.56$ ; action potentials per syllable:  $P = 0.94$ ; paired  $t$ -tests,  $N = 4$  cells, 3 birds). Furthermore, during the period of overlap, neural activity was locked precisely to features of the syllable being sung but not to the playback syllable, even when the two syllables were of the same type (Fig. 4b, d, Supplementary Fig. 2d, e;  $N = 4$  cells, 3 birds). In contrast, recordings made in the absence of singing revealed that auditory activity normally evoked by presentation of the primary song type was strongly attenuated when a phase-delayed copy of the primary song type or another song

was simultaneously presented through another speaker (Fig. 4c–e,  $N = 8$  cells, 4 birds). Together, these observations indicate that singing-related activity in HVC<sub>X</sub> cells is due to corollary discharge



**Figure 3 | HVC<sub>X</sub> neurons exhibit a precise sensorimotor correspondence.** **a, b**, Singing-related and auditory activity (respectively top and middle row) in association with several syllables of the primary song type (shown as a spectrogram, bottom row). **c, d**, Action potential timing was quite similar in the singing and hearing states, both within and across HVC<sub>X</sub> cells. (In **c**,  $P = 0.50$ , paired  $t$ -test. In **d**,  $N = 9$  song types, 7 cells, 3 birds, mean  $\pm$  s.d.; shaded symbols, cells in **a, b**; regression:  $P < 0.01$ ,  $R^2 = 0.99$ , slope = 1.05, intercept =  $-2.90$ ; diagonal line represents identity.) Interestingly, two cells that were active in association with two song types in the bird's repertoire displayed a precise auditory–vocal correspondence for both song types (triangles and squares indicate paired song types; see also Supplementary Fig. 1d).



**Figure 4 | Action potentials in HVC<sub>X</sub> neurons during singing are a corollary discharge of song motor activity.** **a**, Auditory response to the primary song type was suppressed before and during singing (arrows, singing of introductory notes; top, spectrogram of microphone recording; bottom, raw data; see also Supplementary Fig. 2a). **b**, Distorted auditory feedback (DAF; shaded regions) as the bird sang the primary song type did not affect either the probability of occurrence ( $P = 1.00$ ) or the timing of action potentials ( $P = 0.52$ , paired  $t$ -tests; top, syllable raster; second row, PSTH, 5 ms bin size; bottom, spectrogram of the vocalized syllable). **c**, Auditory response (middle) to the primary song type (top) was suppressed when the primary song type was played through a second speaker at a pseudorandom phase delay (bottom). **d**, Acoustic distortion strongly attenuated auditory activity but not singing-related activity (auditory, green,  $P = 0.01$ ,  $N = 8$  cells, 4 birds; singing, red,  $P = 0.59$ ,  $N = 4$  cells, 3 birds). Action potential timing was unaffected by distortion in either state (auditory,  $P = 0.15$ ; singing,  $P = 0.56$ , mean  $\pm$  s.d.; solid lines, control; dotted lines, distortion present). **e**, Acoustic distortion reduced the number of action potentials per syllable in the auditory state (green) but not in the singing (red) state (auditory,  $P = 0.02$ ,  $N = 8$  cells; singing,  $P = 0.93$ , paired  $t$ -tests in all cases;  $N = 4$  cells, mean  $\pm$  s.e.).

rather than an auditory feedback signal and that HVC<sub>X</sub> cells are gated to exist in purely auditory or motor states.

### Sensorimotor correspondence in another species

To investigate whether the sensorimotor correspondence seen in swamp sparrow HVC<sub>X</sub> neurons generalized to HVC<sub>X</sub> cells of other songbirds, we recorded from HVC<sub>X</sub> cells in Bengalese finches (*Lonchura striata domestica*). Adult Bengalese finch song is highly sensitive to distortion of auditory feedback<sup>20,21</sup>, thus affording a more rigorous test of the idea that singing-related activity of HVC<sub>X</sub> cells is due to motor corollary discharge. As observed in swamp sparrows, HVC<sub>X</sub> cells in the awake Bengalese finch responded selectively to playback of the bird's own song (Supplementary Fig. 3a,  $N = 16$  HVC<sub>X</sub> cells, 2 birds). These auditory responses were highly phasic, occurring in association with certain syllables in the song phrase (Supplementary Fig. 3a). In direct parallel with our observations in swamp sparrows, HVC<sub>X</sub> cells in Bengalese finches showed singing-related activity, and auditory and singing-related activities were aligned relative to syllable onset (Supplementary Figs 3a, 4a,  $N = 6$  cells, 2 birds). This singing-related activity was unaffected by distorted auditory feedback (Supplementary Figs 3b, 4a–c,  $N = 5$  cells,

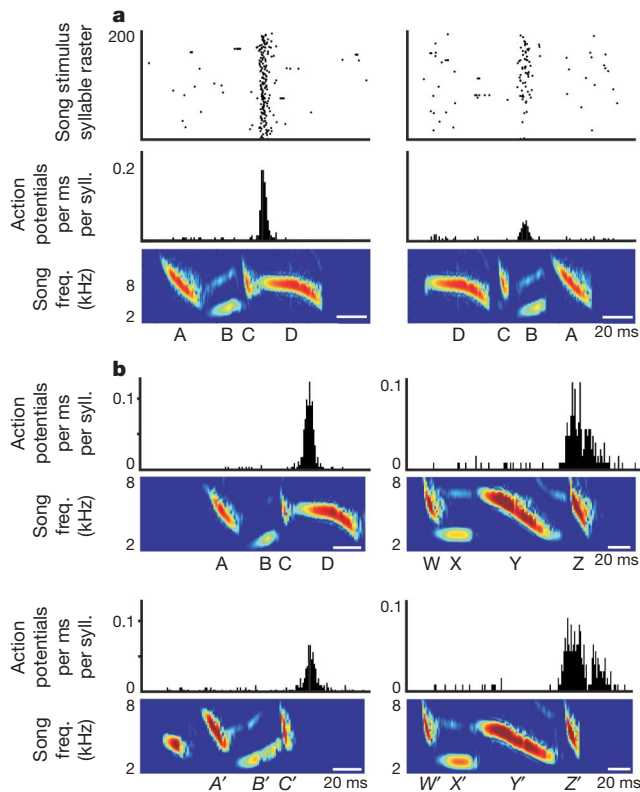
2 birds), while auditory activity was strongly suppressed when two copies of the effective song phrase were presented simultaneously with variable phase offset (Supplementary Fig. 3c,  $N = 6$  cells, 2 birds). Therefore, singing-related activity of HVC<sub>X</sub> neurons is dominated by motor corollary discharge in both swamp sparrows and Bengalese finches. Thus the capacity of HVC<sub>X</sub> cells to exhibit a precise sensorimotor correspondence and switch rapidly between auditory and motor states may constitute a general mechanism underlying learned vocal communication in songbirds.

### Auditory responses extend to other birds' songs

For HVC<sub>X</sub> neurons to facilitate communication, their sensory responsiveness must extend to other birds' songs. In initial experiments in swamp sparrows, we found that HVC<sub>X</sub> cells were unresponsive to other swamp sparrow songs chosen at random (Fig. 1). These conspecific songs may have failed to evoke responses because HVC<sub>X</sub> cells respond exclusively to self-generated vocalizations, or because the songs we chose lacked certain necessary features. Consistent with the idea that HVC<sub>X</sub> cells in swamp sparrows respond to specific features, all HVC<sub>X</sub> cells responded at a precise phase of the syllable presentation. Because note sequences are important features for some auditory HVC neurons<sup>9,22</sup>, we presented artificial trilled syllables containing the primary song type notes in their natural or reverse order (Fig. 5a). Almost all HVC<sub>X</sub> cells tested in this manner (12/14 cells, 7 birds) responded to only the naturally occurring sequence, indicating that a sequence of at least two notes was necessary to elicit an auditory response. We then tested whether HVC<sub>X</sub> neurons would respond to other swamp sparrow songs containing note sequences similar to those in the primary song type. We found that a swamp sparrow song with a note sequence similar to that in the primary song type could drive auditory responses in HVC<sub>X</sub> neurons (Fig. 5b;  $N = 14$  cells including 3 in which both singing and auditory data were collected, 7 birds;  $N = 19$  effective stimuli). In some cases, a conspecific song could evoke a more robust response than that elicited by the primary song type (range: 1.00–1.64 conspecific response normalized to primary song type response;  $N = 5$  cells, 4 birds). When exemplar syllables of the primary song type and an effective song of another sparrow were plotted relative to the average action potential latency for each syllable, the note sequences in the two syllables were aligned (Fig. 5b). Thus, the selective auditory responsiveness of HVC<sub>X</sub> cells extends to similar vocal sequences produced by other birds, making auditory–vocal HVC<sub>X</sub> neurons well suited to a role in communication.

The ability of HVC<sub>X</sub> neurons to respond to other birds' songs and to display an auditory–motor correspondence could facilitate vocal communication in two ways. First, when the sender's vocalizations activate the receiver's auditory–vocal HVC<sub>X</sub> neurons, those vocalizations could be compared to an internal representation of the receiver's vocal gestures, enabling perceptual categorization of songs in the context of the receiver's vocal repertoire<sup>23</sup>. Second, auditory activation of HVC<sub>X</sub> neurons by other birds' songs could provide a template for subsequent movement, enabling the animal to select a vocalization from its repertoire that matches songs of its neighbours. In many regards, auditory–vocal HVC<sub>X</sub> cells are similar to visual–motor 'mirror neurons' in the monkey frontal cortex<sup>24–26</sup> that are hypothesized to play a role in perception of communication gestures<sup>27–29</sup>, including human speech<sup>30,31</sup>. In that light, the precise temporal alignment of auditory and vocal activity in HVC<sub>X</sub> cells suggests that auditory–vocal mirror neurons express an additional mode of sensorimotor correspondence not previously reported for visual–motor mirror neurons. An important remaining question is whether auditory activity in HVC<sub>X</sub> cells is related to the bird's perception of songs, as predicted for mirror neurons.

Beyond serving a perceptual role, auditory–vocal HVC<sub>X</sub> cells could have a role in vocal learning. During singing, HVC<sub>X</sub> neurons transmit song corollary discharge sufficiently delayed to mimic auditory feedback associated with the vocalization. This delay probably arises



**Figure 5 | Swamp sparrow HVC<sub>X</sub> neurons respond to note sequences in the primary song type and to similar note sequences in other swamp sparrows' songs.** **a**, An HVC<sub>X</sub> neuron responded robustly to the primary song type with the notes in the natural sequence (left) but weakly or not at all when the notes were in the reverse order (right). Top row, syllable raster; middle row, PSTH, 1 ms bin size; bottom row, syllable spectrogram. **b**, HVC<sub>X</sub> neurons (left, cell 1; right, cell 2) responded to note sequences in the primary song type (top pair of histogram and spectrogram) and similar sequences in another (conspecific) sparrow's song (bottom pair). Histogram, PSTH, 1 ms bin size; Spectrogram, syllable spectrogram, notes labelled individually. (19 of 23 similar conspecific (CON) songs evoked a response; CON responses normalized to the primary song type response; effective stimuli,  $0.87 \pm 0.32$ ; ineffective stimuli,  $0.28 \pm 0.16$ , mean  $\pm$  s.d.,  $P < 0.01$ , paired  $t$ -test, range of CON responses, 0–1.64; data not shown). Alignment of syllables in the primary song type (top, spectrogram) and effective conspecific song (bottom) using the mean timing of auditory activity revealed similar spectrotemporal features.

when song premotor activity of HVC<sub>RA</sub> cells is relayed by interneurons to HVC<sub>X</sub> cells<sup>32</sup>. Inhibitory interneurons in HVC help shape the temporally precise auditory responses of HVC<sub>X</sub> cells to song<sup>10,33</sup>, suggesting that inhibitory synapses onto HVC<sub>X</sub> cells play an important role in establishing the observed sensorimotor correspondence. In both invertebrates<sup>34</sup> and vertebrates<sup>35</sup>, corollary discharge of central motor commands can serve as an estimate of the anticipated sensory feedback. In HVC, this arrangement could provide a motor-based estimate of auditory feedback<sup>19</sup>, with the useful outcome that differences between the motor estimate and the actual feedback could be used to guide song learning. If this model obtains in songbirds, then HVC<sub>X</sub> cells either transmit estimated feedback to a downstream comparator or are the site of comparison. In support of the idea that the comparator lies downstream of HVC, we observed that the singing-related activity of HVC<sub>X</sub> neurons was insensitive to distorted auditory feedback over acute timescales (Fig. 4); such insensitivity has also been described for HVC<sub>X</sub> cells during juvenile song learning<sup>36</sup>. Alternatively, HVC<sub>X</sub> cells may serve as comparators in which corollary discharge typically overwhelms auditory feedback signals, a mismatch that could facilitate song maintenance. Future studies can determine whether HVC<sub>X</sub> neurons are the site of auditory–vocal comparison by recording from those cells while presenting distorted auditory feedback over a timescale sufficient to induce vocal plasticity.

Finally, because HVC<sub>X</sub> neurons innervate striatal structures<sup>14,37</sup> important for song learning and perception<sup>15,16</sup>, the coding strategy employed by HVC<sub>X</sub> neurons to represent vocal sequences may have implications for learning and perception of speech in humans. In the human brain, cortical neurons similar to HVC<sub>X</sub> auditory–vocal neurons could transmit speech-related auditory and motor information to striatal regions implicated in speech development<sup>38,39</sup>. Furthermore, auditory–vocal mirror neurons with properties similar to the HVC<sub>X</sub> cells described here could bind sensory and motor features of distinct vocal gestures, providing an efficient substrate for rapid decoding and encoding of speech<sup>30,31</sup>.

## METHODS SUMMARY

**Song behaviour.** Birds' song types were recorded in a semi-anechoic chamber, digitized at 25 kHz and saved onto a computer hard drive to be used as stimulus songs. Individual note types were classified by S.P. using established criteria<sup>5</sup>. Similar songs were defined as those containing the same sequence of note categories as in the primary song type. Conspecific songs capable of driving auditory responses expressed a range of spectral similarity to the primary song type, as defined using cross-correlation of the two syllables (correlation value range: 0.17–0.78). Audio files of songs in Figs 1, 2 and 5 are available as Supplementary Information.

**Electrophysiological recordings and analysis.** Individual neurons were recorded extracellularly in awake and freely behaving birds. All HVC neurons from which both auditory and singing data were obtained were identified antidromically using stimulation in area X. Action potentials of individual neurons were discriminated by amplitude (custom software) or on the basis of waveform characteristics (WaveClus<sup>40</sup>), and unit isolation was verified by the presence of a refractory period in the interspike interval histogram. To assess auditory selectivity of an isolated neuron, the bird's own song types and other birds' songs were presented through a loudspeaker in the sound attenuating chamber in which the bird was housed. Singing-related activity was recorded along with the bird's vocalization. Rasters and histograms of action potential activity were constructed by aligning action potentials relative to the beginning of the associated song or syllable. Activity during song presentation or singing was compared against the cell's background firing rate using an activity histogram; any value exceeding the mean background rate plus 5 s.d. was deemed significant. Responses to other birds' songs were normalized to the response to the primary song type, with the criterion for effective stimuli being >0.5.

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

Received 10 October; accepted 19 November 2007.

1. Fee, M. S. & Leonardo, A. Miniature motorized microdrive and commutator system for chronic neural recording in small animals. *J. Neurosci. Methods* **112**, 83–94 (2001).

2. Marler, P. & Tamura, M. Culturally transmitted patterns of vocal behavior in sparrows. *Science* **146**, 1483–1486 (1964).
3. Konishi, M. The role of auditory feedback in the control of vocalization in the white-crowned sparrow. *Z. Tierpsychol.* **22**, 770–783 (1965).
4. Marler, P. & Peters, S. Sparrows learn adult song and more from memory. *Science* **213**, 780–782 (1981).
5. Marler, P. & Pickert, R. Species-universal microstructure in the learned song of the swamp sparrow (*Melospiza georgiana*). *Anim. Behav.* **32**, 673–689 (1984).
6. Nottebohm, F., Stokes, T. M. & Leonard, C. M. Central control of song in the canary, *Serinus canarius*. *J. Comp. Neurol.* **165**, 457–486 (1976).
7. Gentner, T. Q., Hulse, S. H., Bentley, G. E. & Ball, G. F. Individual vocal recognition and the effect of partial lesions to HVC on discrimination, learning, and categorization of conspecific song in adult songbirds. *J. Neurobiol.* **42**, 117–133 (2000).
8. Hahnloser, R. H., Kozhevnikov, A. A. & Fee, M. S. An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* **419**, 65–70 (2002).
9. Margoliash, D. Acoustic parameters underlying the responses of song-specific neurons in the white-crowned sparrow. *J. Neurosci.* **3**, 1039–1057 (1983).
10. Mooney, R. Different subthreshold mechanisms underlie song selectivity in identified HVC neurons of the zebra finch. *J. Neurosci.* **20**, 5420–5436 (2000).
11. Yu, A. C. & Margoliash, D. Temporal hierarchical control of singing in birds. *Science* **273**, 1871–1875 (1996).
12. Mooney, R., Hoese, W. & Nowicki, S. Auditory representation of the vocal repertoire in a songbird with multiple song types. *Proc. Natl Acad. Sci. USA* **98**, 12778–12783 (2001).
13. Wild, J. M., Williams, M. N., Howie, G. J. & Mooney, R. Calcium-binding proteins define interneurons in HVC of the zebra finch (*Taeniopygia guttata*). *J. Comp. Neurol.* **483**, 76–90 (2005).
14. Farries, M. A. & Perkel, D. J. A telencephalic nucleus essential for song learning contains neurons with physiological characteristics of both striatum and globus pallidus. *J. Neurosci.* **22**, 3776–3787 (2002).
15. Scharff, C. & Nottebohm, F. A comparative study of the behavioral deficits following lesions of various parts of the zebra finch song system: implications for vocal learning. *J. Neurosci.* **11**, 2896–2913 (1991).
16. Scharff, C., Nottebohm, F. & Cynx, J. Conspecific and heterospecific song discrimination in male zebra finches with lesions in the anterior forebrain pathway. *J. Neurobiol.* **36**, 81–90 (1998).
17. McCasland, J. S. & Konishi, M. Interaction between auditory and motor activities in an avian song control nucleus. *Proc. Natl Acad. Sci. USA* **78**, 7815–7819 (1981).
18. Dave, A. S. & Margoliash, D. Song replay during sleep and computational rules for sensorimotor vocal learning. *Science* **290**, 812–816 (2000).
19. Troyer, T. W. & Doupe, A. J. An associational model of birdsong sensorimotor learning I. Efference copy and the learning of song syllables. *J. Neurophysiol.* **84**, 1204–1223 (2000).
20. Okanoya, K. & Yamaguchi, A. Adult bengalese finches (*Lonchura striata var domestica*) require real-time auditory feedback to produce normal song syntax. *J. Neurobiol.* **33**, 343–356 (1997).
21. Woolley, S. M. & Rubel, E. W. Bengalese finches *Lonchura striata domestica* depend upon auditory feedback for the maintenance of adult song. *J. Neurosci.* **17**, 6380–6390 (1997).
22. Lewicki, M. S. Intracellular characterization of song-specific neurons in the zebra finch auditory forebrain. *J. Neurosci.* **16**, 5855–5863 (1996).
23. Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. Perception of the speech code. *Psychol. Rev.* **74**, 431–461 (1967).
24. Gallese, V., Fadiga, L., Fogassi, L. & Rizzolatti, G. Action recognition in the premotor cortex. *Brain* **119**, 593–609 (1996).
25. Rizzolatti, G. & Craighero, L. The mirror-neuron system. *Annu. Rev. Neurosci.* **27**, 169–192 (2004).
26. Ferrari, P. F., Gallese, V., Rizzolatti, G. & Fogassi, L. Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *Eur. J. Neurosci.* **17**, 1703–1714 (2003).
27. Iacoboni, M. et al. Grasping the intentions of others with one's own mirror neuron system. *PLoS Biol.* **3**, e79 (2005).
28. Rizzolatti, G., Fogassi, L. & Gallese, V. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Rev. Neurosci.* **2**, 661–670 (2001).
29. Iacoboni, M. et al. Cortical mechanisms of human imitation. *Science* **286**, 2526–2528 (1999).
30. Rizzolatti, G. & Arbib, M. A. Language within our grasp. *Trends Neurosci.* **21**, 188–194 (1998).
31. Arbib, M. A. From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. *Behav. Brain Sci.* **28**, 105–124 (2005).
32. Mooney, R. & Prather, J. F. The HVC microcircuit: the synaptic basis for interactions between song motor and vocal plasticity pathways. *J. Neurosci.* **25**, 1952–1964 (2005).
33. Rosen, M. J. & Mooney, R. Inhibitory and excitatory mechanisms underlying auditory responses to learned vocalizations in the songbird nucleus HVC. *Neuron* **39**, 177–194 (2003).
34. Poulet, J. F. & Hedwig, B. The cellular basis of a corollary discharge. *Science* **311**, 518–522 (2006).
35. Bell, C. C. An efference copy which is modified by reafferent input. *Science* **214**, 450–453 (1981).

36. Kozhevnikov, A. A. & Fee, M. S. Singing-related activity of identified HVC neurons in the zebra finch. *J. Neurophysiol.* **97**, 4271–4283 (2007).
  37. Perkel, D. J., Farries, M. A., Luo, M. & Ding, L. Electrophysiological analysis of a songbird basal ganglia circuit essential for vocal plasticity. *Brain Res. Bull.* **57**, 529–532 (2002).
  38. Vargha-Khadem, F., Gadian, D. G., Copp, A. & Mishkin, M. FOXP2 and the neuroanatomy of speech and language. *Nature Rev. Neurosci.* **6**, 131–138 (2005).
  39. Lai, C. S., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F. & Monaco, A. P. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413**, 519–523 (2001).
  40. Quiroga, R. Q., Nadasdy, Z. & Ben-Shaul, Y. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Comput.* **16**, 1661–1687 (2004).
  41. Hyman, J. Countersinging as a signal of aggression in a territorial songbird. *Anim. Behav.* **65**, 1179–1185 (2003).
- Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).
- Acknowledgements** We thank M. Fee and A. Kozhevnikov for training and assistance in building the miniature microdrives used for these chronic recordings. D. Fitzpatrick, M. Ehlers, M. Platt, H. Greenside and J. Groh provided comments on the manuscript. This work was supported by grants from the NIDCD (R.M.) and the N.S.F. (S.N.). J.P. was supported by an NIH NRSA.
- Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence and requests for materials should be addressed to R.M. ([mooney@neuro.duke.edu](mailto:mooney@neuro.duke.edu)).

## METHODS

**Swamp sparrows.** All procedures were in compliance with recommendations of Duke University Animal Care and Use Committee and state and federal regulations governing the capture and use of wild birds. Birds were caught with mist nets as adults (age > 1 yr) either on winter grounds in Orange County, North Carolina, or on their summer breeding grounds in Crawford County, Pennsylvania. Birds were housed individually throughout their time in the laboratory, both before and during experimentation. Birds were provided with seed and water *ad libitum* and were given a regular supplement of mealworms. Males were identified either by external morphology (breeding season) or by molecular marker techniques<sup>42</sup> (out of season), and females were released. Prior to implantation of the stimulus and recording devices, birds were subjected to gradually lengthening photoperiod (1 h week<sup>-1</sup> from 9:15 up to 15:9 L:D cycle) meant to simulate the onset of the spring breeding season, the only time of year when swamp sparrows sing robustly. This change in photoperiod, combined with a subcutaneous implant of testosterone<sup>43</sup>, was sufficient to induce the birds to sing. Birds were recorded in a semi-anechoic chamber (recorded using a Sony TCM 5000 EV recorder and Shure SM57 microphone), and many examples of song (typically > 100) were recorded from each bird to ensure that the bird's full repertoire was sampled (2 to 5 song types). Although the exact age of birds we used for these experiments was unknown, all songs were crystallized, indicating birds were at least 1 yr old. Exemplars of each song type were digitized (25 kHz) and saved onto a computer hard drive (SIGNAL and LabView software) to be used as stimulus songs. Song stimuli consisted of natural song types and synthetic variants (for example, reverse note order) of those song types from the experimental subject and conspecific birds. Natural song types (unaltered from the original recordings) were used to assess the auditory selectivity of each neuron. Digital editing was used to create synthetic variants of the primary song type in which the notes were arranged in either the natural or reverse order, using the same internote intervals as in the natural song. Copies of this syllable were then concatenated to form songs with the same intersyllable intervals and total song duration as in the natural song.

**Bengalese finches.** Procedures were generally the same as those described for swamp sparrows, except that birds were raised in our aviary (15:9 L:D cycle) housed in communal cages. Subject birds were adult males > 155 days of age; males were distinguished from females by males' expression of song. Because Bengalese finch songs have variable syntax from one song bout to another, several variants of song from the subject bird were used to probe the auditory response of each neuron.

**Microdrive implantation surgery.** Neurons were sampled using a miniaturized micromanipulation device<sup>1</sup> in awake and freely behaving birds. Several days before implantation, birds were transferred from their housing cage to the recording chamber, a sound-attenuating box (Acoustic Systems) where they would reside throughout experimentation. During implantation, adult male swamp sparrows were anaesthetized using isoflurane (inhalation, 1–3% in 100% O<sub>2</sub>) and placed in a stereotaxic device. A small incision was made in the skin overlying the skull, and the outer leaflet of bone was removed over HVC, area X and RA. A small craniotomy (approximately 300 × 300 μm) was made in the inner leaflet over area X, and a small custom-made bipolar stimulus electrode (J.F.P.) was inserted to the proper depth. The implant site was covered with a sterile film and the electrode was secured using dental cement. With the electrode in area X firmly secured, the head was repositioned and the same implant procedure was repeated to place a bipolar stimulus electrode in RA. With both stimulus electrodes firmly in place, another small craniotomy was made directly over HVC. HVC was located by passing brief (~100 μs) current pulses through the stimulating electrodes in area X and RA to generate antidromic activity in HVC, and the boundaries of HVC were defined using a sterilized extracellular electrode (Carbostar 1, Kation Scientific) to observe the extent of the region expressing the resultant antidromic 'hash'. The microdrive recording device was implanted so that the recording electrodes were initially positioned slightly dorsal of HVC. The microdrive was secured to the skull using dental cement (microdrive ~1.2 g including dental cement, birds ~16 g), and the incision site was closed using surgical skin adhesive (Vetbond). The bird was monitored closely until it was fully recovered, typically < 15 min. After the recording session was complete (1–5 weeks), the bird was deeply anaesthetized with equithesin, perfused transcardially with saline and then 4% paraformaldehyde, and the brain was processed histologically. All electrode positions were verified at the end of each experiment using Nissl-stained sagittal sections (thickness 75 μm).

**Experimental protocol.** Birds were allowed to recover for three days following the implantation procedure before recording began. During electrophysiological recording, microdrive electrodes were slowly advanced into HVC while weak electrical stimulation was delivered to the stimulus electrodes in either area X or RA (100 μs pulses, ~100 μA). The boundaries of HVC could be reliably

identified by observing where antidromic activity was evident. Once an electrode was positioned in HVC, the electrode was advanced very slowly so that antidromically-evoked action potentials of individual neurons could be identified. All neural data were amplified, filtered (band pass 500 Hz to 10 kHz), and digitized (25 kHz) to computer file (LabView).

Action potentials of individual units were discriminated using amplitude discrimination of the largest unit in a record (custom software) or discrimination based on waveform characteristics (WaveClus<sup>40</sup>). In both cases, single unit isolation was verified using an interspike interval histogram to test for the presence of a refractory period. Individual units were identified using antidromic stimulation via the electrodes placed in area X and RA or by their characteristic electrophysiological response properties, although all cells from which both auditory and singing data were obtained were identified antidromically. In antidromic identification, HVC<sub>X</sub> units displayed fixed-latency action potential responses to stimulation in area X but no response to stimulation in RA. In contrast, HVC<sub>RA</sub> units displayed fixed-latency action potential responses to stimulation in RA but not in area X. Each of these classes of projection neuron could be distinguished from HVC interneurons, which expressed variable-latency responses to stimulation in either RA<sup>8</sup> or area X and occasionally to stimulation at both sites.

When a single unit had been isolated and identified, song playback of each song type in the bird's repertoire was immediately initiated (10 s quiet interval between each song presentation, stimuli presented in randomized order). Songs were played to the sparrow at 70 dB (peak r.m.s., A-weighted) through a speaker placed 20–35 cm away in the chamber (distance varied according to the bird's location in the cage), and a microphone in the chamber was used to record auditory stimuli and the bird's vocalizations. Playback of the bird's entire song repertoire, as well as songs of conspecific birds and synthetic variants of some of the bird's own song types were used to assess the auditory response of each neuron described in the main text. Auditory responsiveness to songs of other swamp sparrows was assessed using conspecific songs that contained some or all of the same sequence of note types<sup>2</sup> as in the syllable of the corresponding song in the bird's repertoire. Conspecific songs expressed a range of spectral similarity to the repertoire song, as defined using cross-correlation of the two syllables (correlation value range: 0.17–0.78), and all conspecific stimuli were selected before any neural recording.

We enforced the following criteria to qualify a neuron as suitable for further analysis: (1) action potentials must have been reliably distinguishable as belonging to only a single unit, (2) all song types in the bird's repertoire must have been presented as auditory stimuli, and (3) the bird must have sung at least once following implantation of the recording device and stimulus electrodes (this ensured that all birds were in roughly similar behavioural states). Extracellular recordings were collected from 60 individual HVC<sub>X</sub> units (7 birds) and 16 individual HVC<sub>RA</sub> units (5 birds, a subset of the 7 birds in which HVC<sub>X</sub> cells were sampled) that met these criteria.

Singing-related activity of HVC<sub>X</sub> neurons was recorded along with the song itself, using either a voice-triggered recording set-up or by evoking countersinging with song playback (see text). For each bird, these songs were compared against the exemplars recorded before surgery, and in each case we noted that song structure was unchanged, consistent with the crystallized song state. Neural activity associated with singing was recorded and compared against features of the song recorded through the microphone in the recording chamber. As swamp sparrow songs were highly stereotyped from one bout to the next, no time-warping of the data was necessary to permit comparison of data collected during singing and during auditory stimulus presentation. Because Bengalese finch electrophysiological data were compared at the level of one- or two-note sequences, stereotypy on that timescale was sufficiently good that time-warping was also unnecessary for those data. In short, time-warping of the data was not performed in any of the analyses reported here.

**Data analysis.** Action potentials from individual neurons were discriminated and compared against features of either the auditory song stimulus during passive playback or features of the song recorded during singing. Song features were discerned using spectrograms generated in Matlab (Mathworks). All analyses were performed in Matlab using custom software (J.F.P. and Stefan Nenkov).

Rasters and histograms of action potential activity were constructed by aligning discriminated action potentials to the song ('whole-song' analysis, 10 ms bin size in whole-song histograms, for example, Fig. 1). Because swamp sparrow songs consist of trilled syllables separated by brief quiet intervals, an additional technique was possible wherein the onset of each syllable was detected separately and used to align action potentials that occurred in association with each syllable ('single-syllable' analysis, 1 ms histogram bin size in single-syllable histograms, for example, Fig. 3c). In both whole-song and single-syllable analyses, the onset of song during presentation of auditory stimuli was defined as the time that the stimulus presentation began, as recorded in each computer file; onset of song

during singing was computed using the spectrogram of the microphone voltage recorded as the bird sang. The onset of song (or of each syllable following a brief quiet period) was defined as the first time when a song note  $>10$  dB louder than background could be detected. In whole-song analyses, action potential latencies were assigned relative to the onset of the song; in single-syllable analyses, action potential latencies were assigned relative to the onset of each associated song syllable.

In both the whole-song and single-syllable analyses, action potential activity during song presentation or singing was compared against the background firing rate when no stimulus was present, and the mean background rate plus 5 s.d. was taken as the threshold for significance. If the value in any bin in the peri-stimulus time histogram exceeded that threshold (accounting for bin size), the auditory response was deemed significant. In assessment of auditory responses to conspecific songs, responses were normalized using the strength of response to the primary song type in each cell. Normalized responses greater than 0.5 were considered effective stimuli, and responses less than 0.5 were considered ineffective. Results obtained in this manner were in good agreement with visual assessment of the efficacy of an auditory stimulus.

Audio files of songs in Figs 1, 2 and 5 are available as Supplementary Information.

42. Griffiths, R., Double, M., Orr, K. & Dawson, R. A. DNA test to sex most birds. *Mol. Ecol.* **7**, 1071–1075 (1998).
43. Marler, P., Peters, S., Ball, G. F., Dufty, A. M. Jr & Wingfield, J. C. The role of sex steroids in the acquisition and production of birdsong. *Nature* **336**, 770–772 (1988).

# The nonlinear Fano effect

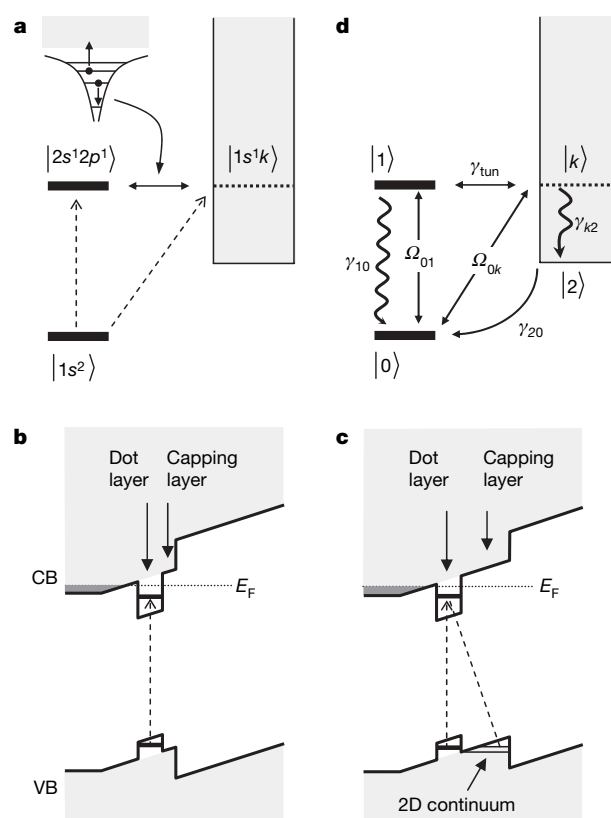
M. Kroner<sup>1\*</sup>, A. O. Govorov<sup>2\*</sup>, S. Remi<sup>1</sup>, B. Biedermann<sup>1</sup>, S. Seidl<sup>1</sup>, A. Badolato<sup>3</sup>, P. M. Petroff<sup>3</sup>, W. Zhang<sup>2</sup>, R. Barbour<sup>4</sup>, B. D. Gerardot<sup>4</sup>, R. J. Warburton<sup>4</sup> & K. Karrai<sup>1</sup>

The Fano effect<sup>1</sup> is ubiquitous in the spectroscopy of, for instance, atoms<sup>1,2</sup>, bulk solids<sup>3,4</sup> and semiconductor heterostructures<sup>5–7</sup>. It arises when quantum interference takes place between two competing optical pathways, one connecting the energy ground state and an excited discrete state, the other connecting the ground state with a continuum of energy states. The nature of the interference changes rapidly as a function of energy, giving rise to characteristically asymmetric lineshapes. The Fano effect is particularly important in the interpretation of electronic transport<sup>5,6</sup> and optical spectra<sup>7,8</sup> in semiconductors. Whereas Fano's original theory<sup>1</sup> applies to the linear regime at low power, at higher power a laser field strongly admixes the states and the physics becomes rich, leading, for example, to a remarkable interplay of coherent nonlinear transitions<sup>9</sup>. Despite the general importance of Fano physics, this nonlinear regime has received very little attention experimentally, presumably because the classic autoionization processes<sup>2</sup>, the original test-bed of Fano's ideas<sup>1</sup>, occur in an inconvenient spectral region, the deep ultraviolet. Here we report experiments that access the nonlinear Fano regime by using semiconductor quantum dots, which allow both the continuum states to be engineered and the energies to be rescaled to the near infrared. We measure the absorption cross-section of a single quantum dot and discover clear Fano resonances that we can tune with the device design or even *in situ* with a voltage bias. In parallel, we develop a nonlinear theory applicable to solid-state systems with fast relaxation of carriers. In the nonlinear regime, the visibility of the Fano quantum interferences increases dramatically, affording a sensitive probe of continuum coupling. This could be a unique method to detect weak couplings of a two-level quantum system (qubits), which should ideally be decoupled from all other states.

We performed our experiments on self-assembled quantum dots. They are known to possess localized discrete energy levels, much like atoms, identified by extremely sharp lines in their optical spectra<sup>10</sup>. Furthermore, when the fundamental cross-gap transition is driven by a strong laser field, the electronic and photon states hybridize. Unmistakable signatures for such dressed states are Rabi oscillations<sup>11,12</sup>, an a.c. Stark effect<sup>10,13</sup>, and a splitting in a resonant high-Q cavity<sup>14–17</sup>. We use InGaAs quantum dots embedded in a GaAs vertical field-effect device<sup>18</sup>. The structure allows us to control the charge stored on an individual quantum dot<sup>18</sup> and to modulate the transition energies by applying a bias voltage, enabling high noise rejection in single dot laser spectroscopy based on modulation techniques<sup>10</sup>.

We present here results for the  $X^{1-}$  exciton transition, which is the transition from a ground state containing a single electron to an excited state containing two electrons and a hole. Sample 1 contains a layer of InGaAs dots separated by a 25 nm tunnel barrier from a GaAs electron reservoir and by a 10 nm capping layer from an AlAs/

GaAs superlattice blocking barrier (Fig. 1b). Laser spectroscopy on dots from this sample shows lorentzian lineshapes (Fig. 2a, b). At low powers, the spectra are independent of power, corresponding to behaviour in the linear regime. At powers above about  $\sim 1$  nW, the spectra depend on power: this is the nonlinear regime. As the power increases, the resonance broadens and the contrast decreases, that is, the resonance saturates, exhibiting power broadening and power-induced transparency<sup>19,20</sup>. The behaviour follows exactly that expected for dressed states in a two-level atom. In sample 2 (Fig. 1c),



**Figure 1 | Schematic level diagrams.** **a**, Classical model of autoionization of a He atom leading to a Fano resonance in absorption. **b**, Level diagram of sample 1, showing the cross-gap exciton transition. **c**, Level diagram of sample 2. In this case, the increased capping layer thickness leads to the appearance of 2D continuum states at the interface between the capping layer and the blocking barrier. These valence continuum states couple via tunnelling with the valence dot level. **d**, Levels, transitions and relaxation processes in the model calculations. CB, conduction band; VB, valence band;  $E_F$ , Fermi energy; see text for definitions of other symbols.

<sup>1</sup>Center for NanoScience and Department für Physik, Ludwig-Maximilians-Universität, 80539 München, Germany. <sup>2</sup>Department of Physics and Astronomy, Ohio University, Athens, Ohio 45701, USA. <sup>3</sup>Materials Department, University of California, Santa Barbara, California 93106, USA. <sup>4</sup>School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK.

\*These authors contributed equally to this work.

the thickness of the capping layer is increased from 10 nm to 30 nm but otherwise the sample was identical to the control, sample 1. In this case, we find that the behaviour at medium and high powers is markedly different: the differential absorption has undershoots and zero crossings (Fig. 2c–h), signatures of Fano-like quantum interferences.

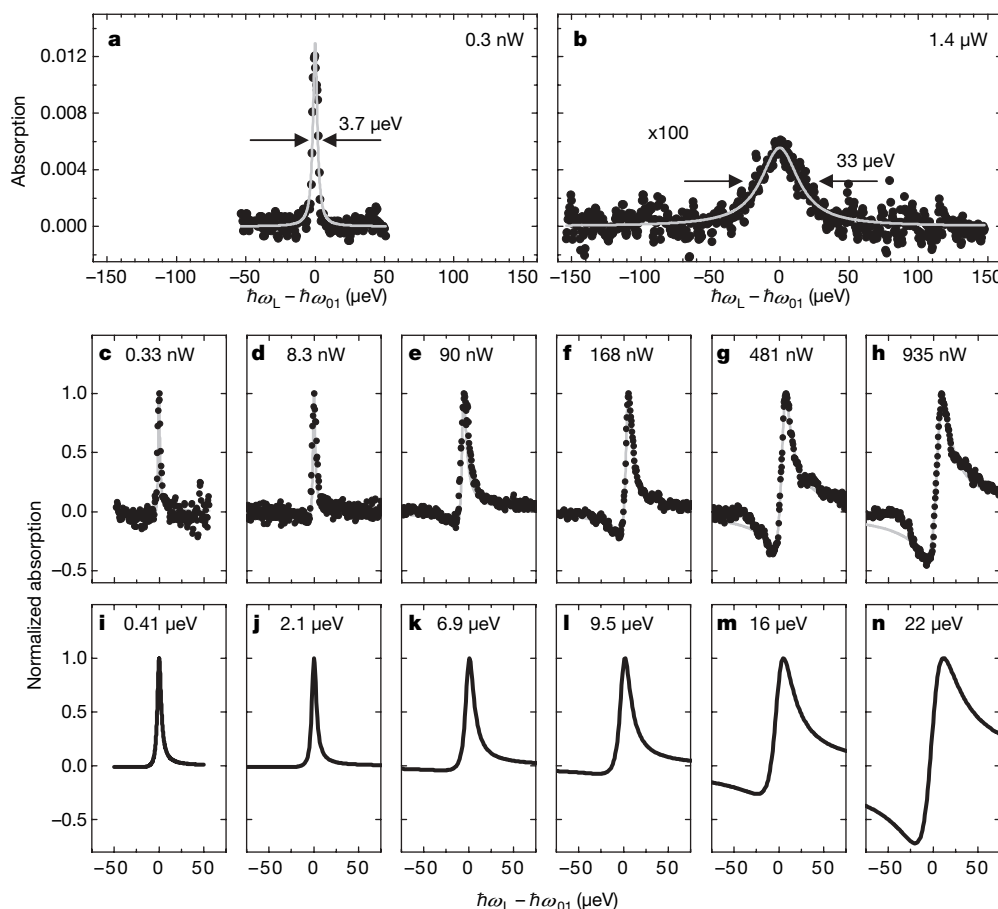
Our key result is that the Fano effects become more and more pronounced as the laser power increases, starting out very small at low power in the linear regime but becoming unmistakable at high power in the nonlinear regime (Fig. 2). The increased visibility of the Fano interference at high laser power results from a different response of the two optical pathways. The optical transition between the discrete levels saturates at high power, but in contrast the weaker continuum transition does not saturate in the range of power we are working in. Increasing the laser power eventually enhances the continuum transition rate to match the saturated discrete level transition. Consequently, the laser power is a convenient experimental control parameter to tune the relative strength of the two competing pathways at the heart of the Fano effect.

This observation, which we back up with the theoretical consideration to follow, represents a highly sensitive technique to detect a very weak coupling between a two-level system and a continuum of extended states when the radiative lifetime of the exciton ( $\tau_{\text{rad}}$ ) is much less than the time required to interact with the continuum (for example, tunnelling or decay time). In the linear regime, the optical

detection of very weak dot–continuum interactions is impossible because the energy uncertainty for the exciton,  $\Delta E$ , obeys the Heisenberg principle  $\Delta E \tau_{\text{rad}} \approx \hbar$ .  $\Delta E$  is equivalently the broadening of the exciton line. In other words, a strong broadening ( $\Delta E \approx \hbar/\tau_{\text{rad}}$ ) makes the dot–continuum interaction invisible in the absorption spectrum. But in the nonlinear regime, the radiative broadening does not play the leading role, and even a very weak dot–continuum interaction becomes apparent (as shown in Fig. 2).

To verify the asymmetries shown in Fig. 2c–h as Fano interferences, we present in Fig. 3b the voltage dependence, monitoring the strength of the interference with the asymmetry parameter  $1/q$ , where  $q$  is the Fano factor determined at constant power. In its standard definition,  $q$  is infinite when the continuum transition is very weak, in which case the line shape is symmetric and entirely determined by the discrete transition. In contrast, when  $q$  is near unity, both the continuum and discrete optical transitions are of similar strengths, and the line shape becomes very asymmetric. The  $1/q$  parameter has a strong bias dependence, disappearing towards the right-hand edge of the  $X^{1-}$  plateau. The bias dependence cannot be explained by a purely optical interference<sup>21,22</sup>, as in this case the bias would have no effect. Instead, a Fano interference provides a natural explanation.

Optical excitation drives the system from its discrete initial state, a quantum dot containing a single electron,  $|0\rangle$ , to the  $X^{1-}$  state containing two electrons and one hole,  $|1\rangle$  (Fig. 1d). State  $|1\rangle$  is in tunnel



**Figure 2 | Laser spectroscopy on a single quantum dot.** **a, b**, Absorption of a single quantum dot from sample 1, exhibiting two-level behaviour, plotted against detuning for two laser intensities, 0.3 nW in the linear regime (**a**), and 1.4 μW in the nonlinear regime (**b**). The solid lines are lorentzian fits to the data. The observed nonlinear behaviour indicates that the dot has no significant coupling to extended electronic states of the crystal. **c–h**,  $X^{1-}$  absorption spectra from a single dot from sample 2 for several different laser powers as indicated in the figure; the absorption spectrum is given by the

change in transmission  $\Delta T(\delta)/T$ , where  $\delta = \omega_L - \omega_{01}$  is the detuning and  $T$  is the transmission. Symbols represent the experiment, solid lines are a guide to the eye based on Fano's theory. **i–n**, Absorption spectra as calculated with the theory described in the text with parameters:  $\hbar\gamma_{10} = 2.16$  μeV,  $q = 12$ ,  $\Delta = 0.4$  μeV and  $\hbar\gamma_{20} = 30$  μeV. The Rabi energies ( $\hbar\Omega_{01}$ ) indicated in the panels correspond to the laser powers of the experiment. The data were measured at 4.2 K with a wavelength of ~950 nm on the  $X^{1-}$  resonance.

contact with the continuum: the combination of applied electric field and large capping thickness enables the hole to tunnel out of the dot into an empty continuum state<sup>23</sup>,  $|k\rangle$  (Fig. 1c). The final state of the transition is therefore hybridized with the continuum. Furthermore, a weak optical transition must also exist between  $|0\rangle$  and  $|k\rangle$ . The two conditions for the Fano effect—two competing optical pathways and a hybridized excited state—are satisfied. The tunnelling involves a bound hole and the valence states at the capping layer–blocking barrier interface, which have a two-dimensional (2D) character. The tunnelling rate is non-zero when the localized hole level is within the 2D continuum of hole states in the quantum well<sup>23</sup>. This is the explanation for the bias dependence in Fig. 3b: at gate voltage  $V_g > -0.22$  V, the quantum dot state moves out of the 2D continuum, the hybridization with the continuum vanishes and  $1/q \rightarrow 0$ . The modelling of  $1/q(V_g)$  confirms our picture of tunnelling (Fig. 3b and Supplementary Materials). We stress that virtually any dot in sample 2 shows a nearly symmetric line at low power with quite large  $q$  ( $\sim 12$ ), and strongly asymmetric Fano lines at elevated powers.

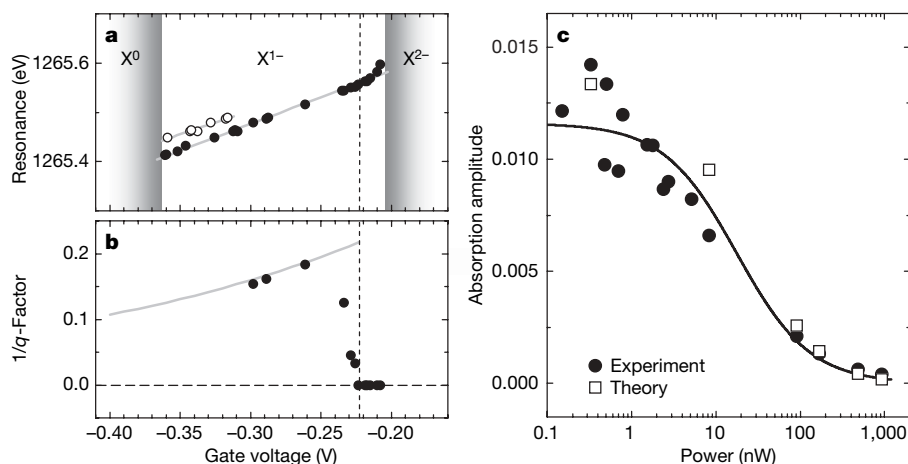
We present a quantum mechanical model of these processes. Fano's original theory<sup>1</sup> applies for a weak driving field where hybridization arises between basis states  $|1\rangle$  and  $|k\rangle$ . The classic case is the doubly excited state of the He atom. This state can auto-ionize (Fig. 1a). In the presence of a strong driving field, the mathematics is doubly difficult but nevertheless analytical results for atomic systems exist<sup>9,24</sup>. However, the solid-state systems are very different and require a new theory. First, in our case, the quantum dot ground state  $|0\rangle$  is repopulated through tunnelling from the reservoir. In the atomic case, the Fano resonance leads to photoionization—excited electrons are ejected at high speed and never return. The large time limits are therefore different: in the quantum dot case, the steady state corresponds to non-zero absorption; in the atomic case it does not, as the system becomes ionized and absorption vanishes. Second, energy relaxation processes are crucial in the quantum dot system but not in the atomic system. For instance, the autoionization rate of the  $2s^1 2p^1$  He atom state  $|1\rangle \rightarrow |k\rangle$  is much faster than spontaneous emission and  $q \approx 1$ , whereas in our sample 2, the tunnelling rate,  $\Delta/\hbar$ , is less than the spontaneous emission rate,  $\Delta/\hbar \leq \gamma_{10}$ , and  $q \gg 1$ ; here  $\Delta$  and  $\gamma_{10}$  represent the tunnel broadening and spontaneous emission rate, respectively. In this sense, quantum dot quantum optics can be very different to atom optics as the parameters and conditions can be widely different and controllably designed.

We present a generalized theory for a closed system. The levels in our model are sketched in Fig. 1d: a ground state,  $|0\rangle$ , a single exciton state  $|1\rangle$ , and a set of continuum states,  $|k\rangle$ . Optical absorption processes are described by the matrix elements (Rabi frequencies)  $\Omega_{01}$  and  $\Omega_{0k}$ . States  $|1\rangle$  and  $|k\rangle$  are connected by a tunnel matrix element,  $\gamma_{\text{tun}}$ . Fast relaxation in the continuum states is described with a relaxation rate  $\gamma_{k2}$  from state  $|k\rangle$  to a shelving state  $|2\rangle$ ; repopulation from  $|2\rangle$  to  $|0\rangle$  is described by rate  $\gamma_{20}$ . In practice, hole tunnelling leaves behind two electrons in the dot. The hole relaxes rapidly on a picosecond timescale to the bottom of the 2D hole continuum; of the two electrons in the dot, one tunnels out to the back contact on a timescale of  $\sim 10$  ps (ref. 25) and it is this tunnelling process that is described with rate  $\gamma_{20}$ . The  $X^{1-}$  spectra are recorded in the voltage interval where the electron state is stable owing to the strong Coulomb blockade, allowing us to neglect Kondo-like processes<sup>26</sup>. In order to include the various relaxation processes, we perform the calculation with a master equation for the density matrix. We obtain an analytic result for the optical absorbed power,  $Q(\delta)$  (where  $\delta$  is the detuning of the laser from the resonance; see below), under the realistic assumption of a small population of the continuum states; indeed, filling of the states in the continuum would require very large powers. The resulting equation is rather complex (equation (S2), Supplementary Information) but yields the correct limit in the linear regime,  $\Omega_{01} \ll \gamma_{10}$ :

$$Q(\delta) = \hbar^2 \Omega_{01}^2 \frac{\omega_L}{2q^2 \Delta} \left( 1 + \frac{\gamma(q^2 - 1)(\Delta/\hbar)}{\gamma^2 + \delta^2} + \frac{2q\delta(\Delta/\hbar)}{\gamma^2 + \delta^2} \right) \quad (1)$$

where  $\delta = \omega_L - \omega_{01}$ , the detuning of the laser ( $\hbar\omega_L$ ) from the resonance ( $\hbar\omega_{01}$ );  $\Delta = \pi \hbar^2 \gamma_{\text{tun}}^2 \rho$ , the tunnel broadening where  $\rho$  is the density of continuum states;  $q = \hbar \gamma_{\text{tun}} \Omega_{01} / \Delta \Omega_{0k}$  the Fano factor; and  $\gamma = \gamma_{10} + \Delta/\hbar$ . We assume here that the effective width of the 2D continuum is much larger than  $\Omega_{01}$  and  $\Delta$ . This is similar to saying that  $\gamma_{\text{tun}}$  is a slowly varying function of  $k$  and gate voltage (see Supplementary Materials). The final term in equation (1) is negligible when  $q \gg 1$  and  $\Delta/\hbar < \gamma_{10}$ , leading to an almost symmetric lineshape. In other words, the tunnel coupling is masked by strong radiative damping. In the limit  $\Omega_{01} \gg \Delta/\hbar, \gamma_{10}$ , the line becomes symmetric but instead of a maximum it has a minimum at  $\delta = 0$ :

$$Q(\delta) = \hbar^2 \Omega_{01}^2 \frac{\omega_L}{2q^2 \Delta} \left( 1 - \frac{(q^2 + 1)\Omega_{01}^2}{4q^2 \delta^2 + 2(q^2 + 1)\Omega_{01}^2} \right) \quad (2)$$



**Figure 3 | Voltage dependence of the Fano resonance.** **a**, Energy of the  $X^{1-}$  resonance ( $\hbar\omega_{01}$ ) as a function of gate voltage (filled circles). The  $X^{1-}$  is observed in the window of gate voltages between  $-0.36$  V and  $-0.205$  V as a result of Coulomb blockade. At the low energy side of the plateau a second resonance line appears, as depicted by the open circles. The vertical dashed line shows the onset of the asymmetry, and the solid line represents the linear Stark effect. **b**,  $1/q$  versus gate voltage at a constant power of 200 nW.

At gate voltages below  $-0.3$  V, the second peak hinders the fitting of the spectra. The solid line reflects the calculated voltage dependence of  $1/q$  (see Supplementary Materials). The horizontal dashed line represents the zero level. **c**, The measured absorption amplitude of the resonance as a function of laser power is plotted, as well as the amplitudes predicted by the theory with the parameters in Fig. 2. The line represents a two-level model as a guide to the eye.

This behaviour corresponds to a 'negative resonance' or strongly destructive interference in the limit of very large power. At realistic, finite  $\Omega_{01}$ , the destructive interference shows up as a large undershoot to the resonance with a zero crossing (Fig. 2n).

The experimental data in Fig. 2 are given for the absorption coefficient  $\alpha = \Delta T/T = Q(\delta)/P$ , where  $T$  and  $P$  are the transmission and light power, respectively. In the linear regime, the maximum absorption coefficient depends on the laser spot area  $A$  as  $\alpha_0 = 3(\lambda/n)^2/(2\pi A)$ , where  $\lambda$  and  $n$  are the laser wavelength and material refractive index, respectively. This expression for  $\alpha_0$  follows from  $\Omega_{01}^2 = 2\alpha_0\gamma_{10}P/(\hbar\omega_L)$  (ref. 20). We compare the theory to the experimental data by taking the known values of  $\gamma_{10}$ , by determining  $q$  through a fit to the data at low power, and then adjusting  $\Delta$  and  $\gamma_{20}$  to account for the data at high power. We find very good agreement with the theory (Fig. 2i–n and Fig. 3c). Here  $\Delta$  is small because the tunnel coupling is weak. Simultaneously,  $q$  is large because the inter-band optical element  $\Omega_{0k}$  is small.

The significance of the negative signals in Fig. 2 is that even very weak coupling to the continuum becomes easy to detect by enhancing the interference at large laser power. At small power, the fundamental spontaneous emission process destroys the interference effect. In principle, the spontaneous emission could be suppressed in a detuned microcavity<sup>16,27</sup>. However, this method is very challenging technologically, and would require elaborate sample preparations. Our method is certainly more flexible. In the control sample 1, it now becomes striking that there are no hints of the Fano effect even at the highest power, demonstrating that in this case, the quantum dot behaves very much like a few-level system. We should note that, with the nonlinear Fano effect, we are able to suppress the role of spontaneous emission dephasing in our quantum dots; however, other types of dephasing need to be analysed specially and, in principle, they may wash out the Fano asymmetry. Fortunately, the exciton resonance in our dots is predominantly dephased by spontaneous emission<sup>10</sup>.

Two overriding points emerge. The first is the tunability of the quantum dot system: Fano effects can be turned on and off. The second is that the nonlinear Fano effect can be used to detect very weak interactions with continuum states in quantum systems. The nature of the interaction is not restricted: tunnelling, Auger processes and Foerster transfer are all included<sup>28</sup>. We note that a very strong nonlinear Fano effect was also observed on p-doped samples, in which the continuum of states is most probably generated by impurity states due to the doping atoms. The nonlinear Fano resonance described here could be produced by interactions of many different types—this is because the three-state scheme demonstrating the quantum interference effect (Fig. 1 c) is generic, and appears in a variety of physical systems, including solids, atoms, molecules and photonics.

Received 7 August; accepted 22 November 2007.

1. Fano, U. Effects of configuration interactions on intensities and phase shifts. *Phys. Rev.* **124**, 1866–1878 (1961).
2. Madden, R. P. & Codling, K. New autoionizing atomic energy levels in He, Ne, and Ar. *Phys. Rev. Lett.* **10**, 516–518 (1963).
3. Cerdeira, F., Fjeldly, T. A. & Cardona, M. Effect of free carriers on zone-center vibrational modes in heavily doped p-type Si. II. Optical modes. *Phys. Rev. B* **8**, 4734–4745 (1973).

4. Hase, M., Demsar, J. & Kitajima, M. Photoinduced Fano resonance of coherent phonons in zinc. *Phys. Rev. B* **74**, 212301 (2006).
5. Faist, J., Capasso, F., Sirtori, C., West, K. W. & Pfeiffer, L. N. Controlling the sign of quantum interference by tunnelling from quantum wells. *Nature* **390**, 589–592 (1997).
6. Schmidt, H., Campman, K. L., Gossard, A. C. & Imamoglu, A. Tunneling induced transparency: Fano interference in intersubband transitions. *Appl. Phys. Lett.* **70**, 3455–3457 (1997).
7. Bar-Ad, S., Kner, S., Marquinezini, M. V., Mukamel, S. & Chemla, D. S. Quantum confined Fano interference. *Phys. Rev. Lett.* **78**, 1363–1366 (1997).
8. Wagner, J. & Cardona, M. Electronic Raman scattering in heavily doped p-type germanium. *Phys. Rev. B* **32**, 8071–8077 (1985).
9. Rzaewski, K. & Eberly, J. H. Confluence of bound-free coherences in laser-induced autoionization. *Phys. Rev. Lett.* **47**, 408–412 (1981).
10. Högele, A. *et al.* Voltage-controlled optics of a quantum dot. *Phys. Rev. Lett.* **93**, 217401 (2004).
11. Zrenner, A. *et al.* Coherent properties of a two-level system based on a quantum dot photodiode. *Nature* **418**, 612–614 (2002).
12. Gammon, D. & Steel, D. G. Optical studies of single quantum dots. *Phys. Today* **55**, 36–41 (2002).
13. Stuffer, S., Ester, P., Zrenner, A. & Bichler, M. Quantum optical properties of a single  $\text{In}_{0.5}\text{Ga}_{0.5}\text{As}$ -GaAs quantum dot two-level system. *Phys. Rev. B* **72**, 121301(R) (2005).
14. Reithmaier, J. P. *et al.* Strong coupling in a single quantum dot-semiconductor microcavity system. *Nature* **432**, 197–200 (2004).
15. Yoshie, T. *et al.* Vacuum Rabi splitting with a single quantum dot in a photonic crystal nanocavity. *Nature* **432**, 200–203 (2004).
16. Peter, E. *et al.* Exciton-photon strong-coupling regime for a single quantum dot embedded in a microcavity. *Phys. Rev. Lett.* **95**, 067401 (2005).
17. Hennessy, K. *et al.* Quantum nature of a strongly coupled single quantum dot-cavity system. *Nature* **445**, 896–899 (2007).
18. Warburton, R. J. *et al.* Optical emission from a charge-tunable quantum ring. *Nature* **405**, 926–929 (2000).
19. Loudon, R. *The Quantum Theory of Light* 3rd edn (Oxford Univ. Press, Oxford, 2000).
20. Kroner, M. *et al.* Resonant saturation laser spectroscopy of a single self-assembled quantum dot. *Physica E* (in the press).
21. Alén, B. *et al.* Absorptive and dispersive optical responses of excitons in a single quantum dot. *Appl. Phys. Lett.* **89**, 123124 (2006).
22. Atatüre, M. *et al.* Observation of Faraday rotation from a single confined spin. *Nature Phys.* **3**, 101–105 (2007).
23. Seidl, S. *et al.* Absorption and photoluminescence spectroscopy on a single self-assembled charge tunable quantum dot. *Phys. Rev. B* **72**, 195339 (2005).
24. Rzaewski, K. & Eberly, J. H. Photoexcitation of an autoionizing resonance in the presence of offdiagonal relaxation. *Phys. Rev. A* **27**, 2026–2042 (1983).
25. Smith, J. M. *et al.* Voltage control of the spin dynamics of an exciton in a semiconductor quantum dot. *Phys. Rev. Lett.* **94**, 197402 (2005).
26. Govorov, A. O., Warburton, R. J. & Karrai, K. Kondo excitons in self-assembled quantum dots. *Phys. Rev. B* **67**, 241307(R) (2003).
27. Bayer, M. *et al.* Inhibition and enhancement of the spontaneous emission of quantum dots in structured microresonators. *Phys. Rev. Lett.* **86**, 3168–3171 (2001).
28. Zhang, W., Govorov, A. O. & Bryant, G. W. Semiconductor-metal nanoparticle molecules: hybrid excitons and non-linear Fano effect. *Phys. Rev. Lett.* **97**, 146804 (2006).

Supplementary Information is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank A. Högele for discussions and J. P. Kotthaus for support. The work was supported by SFB 631 (Germany), AvHF (Germany), EPSRC (UK), NSF (USA) and SANDIE (EU). B.D.G. thanks the Royal Society of Edinburgh for financial support. Financial support from the German Excellence Initiative via the Nanosystems Initiative Munich (NIM), and from Ohio University Nanobiotechnology Initiative, is acknowledged.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence and requests for materials should be addressed to A.O.G. ([govorov@helios.phy.ohiou.edu](mailto:govorov@helios.phy.ohiou.edu)).

# Reduction and selective oxo group silylation of the uranyl dication

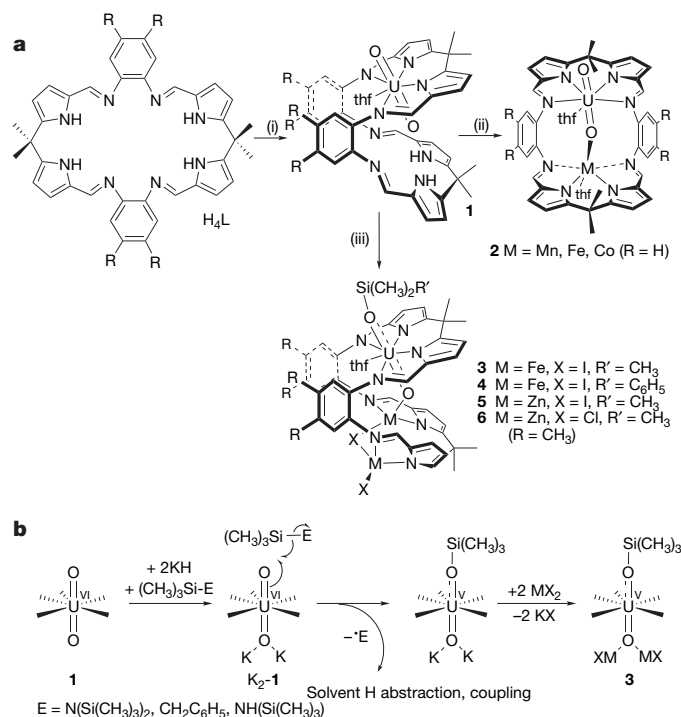
Polly L. Arnold<sup>1</sup>, Dipti Patel<sup>1</sup>, Claire Wilson<sup>2</sup> & Jason B. Love<sup>1</sup>

Uranium occurs in the environment predominantly as the uranyl dication  $[\text{UO}_2]^{2+}$ . Its solubility renders this species a problematic contaminant<sup>1–3</sup> which is, moreover, chemically extraordinarily robust owing to strongly covalent U–O bonds<sup>4</sup>. This feature manifests itself in the uranyl dication showing little propensity to partake in the many oxo group functionalizations and redox reactions typically seen with  $[\text{CrO}_2]^{2+}$ ,  $[\text{MoO}_2]^{2+}$  and other transition metal analogues<sup>5–9</sup>. As a result, only a few examples of  $[\text{UO}_2]^{2+}$  with functionalized oxo groups are known. Similarly, it is only very recently that the isolation and characterization of the singly reduced, pentavalent uranyl cation  $[\text{UO}_2]^+$  has been reported<sup>10–12</sup>. Here we show that placing the uranyl dication within a rigid and well-defined molecular framework while keeping the environment anaerobic allows simultaneous single-electron reduction and selective covalent bond formation at one of the two uranyl oxo groups. The product of this reaction is a pentavalent and monofunctionalized  $[\text{O}=\text{U} \cdots \text{OR}]^+$  cation that can be isolated in the presence of transition metal cations. This finding demonstrates that under appropriate reaction conditions, the uranyl oxo group will readily undergo radical reactions commonly associated only with transition metal oxo groups. We expect that this work might also prove useful in probing the chemistry of the related but highly radioactive plutonyl and neptunyl analogues found in nuclear waste.

Reactions of the uranyl dication that result in the functionalization or transformation of the U=O groups are rare. Examples include atypical Lewis base behaviour of the uranyl dioxo group towards alkali metals in the solid state<sup>13,14</sup>, and the formation of an unusual  $\text{O}=\text{U} \cdots \text{B}(\text{C}_6\text{F}_5)_3$  adduct involving significant and asymmetric U=O bond lengthening<sup>15</sup>. Photolysis of uranyl phosphine oxide complexes in the presence of alcohols results in two-electron reduction and the formation of U(IV) alkoxides, via the highly oxidizing  $^*\text{UO}_2^{2+}$  excited state; the U(IV) complexes can be hydrolysed to regenerate the uranyl dication cleanly<sup>16</sup>. Usually, the  $[\text{UO}_2]^{2+}$  cation spontaneously disproportionates to  $[\text{UO}_2]^{2+}$  and U(IV) phases in an aqueous environment. We reported recently<sup>17</sup> that the reaction between the mono-uranyl complex, **1** ( $\text{R} = \text{H}$ ), and transition metal silylamides  $[\text{M}\{\text{N}(\text{Si}(\text{CH}_3)_2)_2\}_2]$  ( $\text{M} = \text{Mn, Fe, Co}$ ) forms the molecular cation–anion complexes, **2**, in which, uniquely, the transition metal bonds to the *endo*-uranyl oxygen atom (Fig. 1a), that is, the uranyl acts as a Lewis base to the transition metal<sup>18</sup>; in this case, no electron transfer between the metals was seen. In search of alternative synthetic routes, we have found that the one-pot reaction between **1** ( $\text{R} = \text{CH}_3$ ),  $\text{FeI}_2$ , and the silylamide base  $\text{KN}(\text{Si}(\text{CH}_3)_2)_2$  at  $-78^\circ\text{C}$  resulted in the formation of the new cation–anion complex  $[\text{UO}(\text{OSi}(\text{CH}_3)_3)(\text{thf})\text{Fe}_2\text{I}_2(\text{L})]$ , **3**, in 80% isolated yield, Fig. 1a (see Methods and Supplementary Information for synthetic details; thf stands for tetrahydrofuran).

The X-ray single-crystal structure of **3** (Fig. 2a, and Supplementary Information) shows that the macrocycle geometry remains wedge-shaped, even though two tetrahedral Fe cations are now incorporated

in the lower cavity, and a  $\text{Si}(\text{CH}_3)_3$  group is bound to the *exo*-uranyl oxygen. The uranyl cation displays a distorted pentagonal bipyramidal geometry with a linear  $\text{O1}–\text{U1}–\text{O2}$  group ( $172.16(17)^\circ$ ). The U–O bond distances confirm that the uranyl fragment in **3** is in the pentavalent oxidation state. The *endo*-U1–O1 ( $1.870(4) \text{ \AA}$ ) bond distance in **3** is elongated compared with those of the hexavalent  $[\text{UO}_2]^{2+}$  complexes **1** ( $\text{R} = \text{H}$ : U1–O1  $1.790(4) \text{ \AA}$ ) and **2** ( $\text{M} = \text{Mn}$ : U1–O1  $1.808(4) \text{ \AA}$ ), and is similar to experimental<sup>10–12,19</sup> and calculated<sup>20,21</sup> bond distances for pentavalent  $[\text{UO}_2]^+$  (range  $1.811$  to  $1.934 \text{ \AA}$ ). The *exo*-U1–O2 ( $1.993(4) \text{ \AA}$ ) bond distance is appreciably longer than U1–O1 (compare with **2** ( $\text{M} = \text{Mn}$ ): U1–O2  $1.768(5) \text{ \AA}$ ), but is significantly shorter than in tetravalent U–OSiR<sub>3</sub> complexes<sup>22</sup> and pentavalent U–OR compounds<sup>23</sup> (all greater than  $2.0 \text{ \AA}$ ). This implies that the *exo*-U–O bond still retains some multiple bond character, but less than that of the *endo*-U–O bond. Both Fe1 and Fe2 are four-coordinate and bound to the macrocycle by single iminopyrrolides, and to each other



**Figure 1 | Reductive silylation of the uranyl dication.** **a**, Synthesis of the uranyl complex **1** and cation–anion complexes. **b**, Proposed mechanism. Reagents and conditions (i)  $[\text{UO}_2(\text{thf})_2\{\text{N}(\text{Si}(\text{CH}_3)_2)_2\}_2]$ , thf (thf = tetrahydrofuran); (ii)  $[\text{M}\{\text{N}(\text{Si}(\text{CH}_3)_2)_2\}_2]$ , thf, heat ( $\text{M} = \text{Mn, Fe, Co}$ ;  $\text{R} = \text{H}$ ); (iii) either  $\text{KN}(\text{Si}(\text{CH}_3)_2)_2$ ,  $\text{MX}_2$  ( $\text{M} = \text{Fe, X} = \text{I, R}' = \text{CH}_3, \text{C}_6\text{H}_5$ ;  $\text{M} = \text{Zn, X} = \text{Cl, I, R}' = \text{CH}_3$ ) or  $\text{KH, FeI}_2, \text{N}(\text{Si}(\text{CH}_3)_3)_3$  or  $\text{C}_6\text{H}_5\text{CH}_2\text{Si}(\text{CH}_3)_3$ ; thf,  $-80^\circ\text{C}$ ,  $\text{R} = \text{CH}_3$ .

<sup>1</sup>School of Chemistry, University of Edinburgh, West Mains Road, Edinburgh EH9 3JJ, UK. <sup>2</sup>Rigaku Europe, Chaucer Business Park, Watery Lane, Sevenoaks, Kent TN15 6QY, UK.

by a bridging iodide (Fe1–I1 2.7317(13) Å, Fe2–I1 2.6335(13) Å, Fe1–I1–Fe2 70.30(3)°). Notably, Fe1 bonds to the *endo*-uranyl oxygen (Fe1–O1 1.946(4) Å) at a distance commensurate with a single dative bond. The Fe-bridging iodide refined to 79.7(3)% occupancy; after exploration of a number of alternative models the remaining electron density was best modelled as a bridging chloride, considering both the quality of the refinement and comparison of the resulting geometry with literature values. The chloride contaminant has accumulated in the crystal, and derives from amounts present in the original  $[\text{UO}_2(\text{thf})_2\{\text{N}(\text{Si}(\text{CH}_3)_3)_2\}_2]$  starting material.

We carried out experiments to probe the origin of the  $\text{Si}(\text{CH}_3)_3$  group and to confirm the single electron transfer to form pentavalent uranyl. A mixture of **1**,  $\text{FeI}_2$ , and the phenyl-substituted  $\text{KN}(\text{Si}(\text{CH}_3)_2\text{C}_6\text{H}_5)_2$  reacts to afford the phenylsilyl-functionalized  $[\text{UO}(\text{OSi}(\text{CH}_3)_2\text{C}_6\text{H}_5)(\text{thf})\text{FeI}_2(\text{L})]$  **4**, in high yield (see Supplementary Information). Thus, it is clear that the silyl group originates from either the silylamide base,  $\text{KN}(\text{Si}(\text{CH}_3)_2\text{R}')_2$ , or its by-product, the disilazane  $\text{HN}(\text{Si}(\text{CH}_3)_2\text{R}')_2$  ( $\text{R}' = \text{CH}_3, \text{C}_6\text{H}_5$ ). Analysis of the mass balance for the by-product KI shows that two molar equivalents are formed during the reaction, which implies that electron transfer from  $\text{KN}(\text{Si}(\text{CH}_3)_2\text{R}')_2$  does not occur; that is, the silylamide acts solely as a base, and the  $\text{HN}(\text{Si}(\text{CH}_3)_2\text{R}')_2$  by-product formed during the reaction provides the silyl group. In contrast, chemical analogues from the same group as uranium, the molybdenum and tungsten *cis*-dioxo complexes  $[\text{M}^{\text{VI}}\text{O}_2(\text{L}')_2]^{2-}$  ( $\text{M} = \text{Mo}, \text{W}$ ;  $\text{L}' = 1,2\text{-S}_2\text{C}_6\text{H}_4$ ), are readily silylated, even in the absence of redox reactions, to afford  $[\text{M}^{\text{VI}}\text{O}(\text{OSi}(\text{C}_6\text{H}_5)_2(\text{C}_4\text{H}_9))(\text{L}')_2]^{2-}$ . Furthermore, the silylated Mo compound is rapidly hydrolysed to the Mo(IV) mono-oxo compound  $[\text{Mo}^{\text{IV}}\text{O}(\text{L}')_2]^{2-}$  (refs 24,25).

The isolation of the closed-shell Zn(II) compounds **5** and **6** confirms that the transition metal simply stabilizes the pentavalent  $[\text{UO}(\text{OSi}(\text{CH}_3)_2\text{R}')^+]$  fragment, without participating in redox chemistry. Reaction between **1**,  $\text{KN}(\text{Si}(\text{CH}_3)_3)_2$ , and  $\text{ZnX}_2$  ( $\text{X} = \text{Cl}, \text{I}$ ) resulted in the formation of orange/brown  $[\text{UO}(\text{OSi}(\text{CH}_3)_3)(\text{thf})\text{Zn}_2\text{X}_2(\text{L})]$ , ( $\text{X} = \text{I}$ ; **5**,  $\text{Cl}$ ; **6**), in moderate yields, Fig. 1a (see online Methods and Supplementary Information). The X-ray crystal structure of **5** (Fig. 2b, and Supplementary Information) is similar to that of **3**, again with trace chloride incorporated but in this case with an occupancy of 52.7(3)%. The  $\text{U}\cdots\text{O}$  bond distances in **5** (U1–O1 1.867(3) Å, U1–O2 1.975(3) Å) are similar to those in **3**, and are also consistent with pentavalent uranyl. The  $\text{U}=\text{O}$  asymmetric stretch in the infrared spectra of uranyl compounds is normally diagnostic, and should decrease by 100–180  $\text{cm}^{-1}$  on reduction to  $[\text{UO}_2]^+$  (ref. 12).

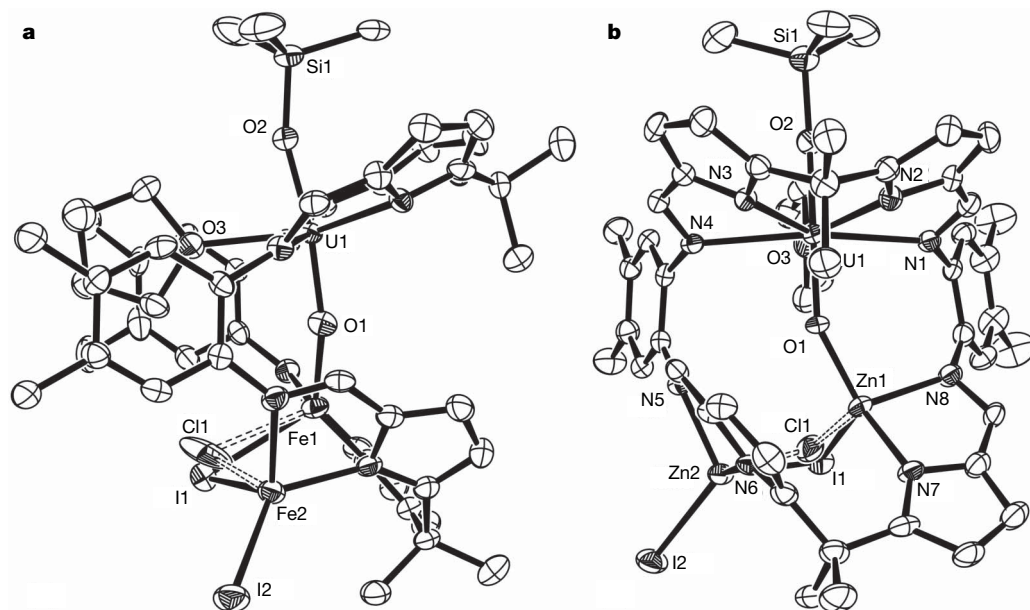
However, the infrared spectra of pentavalent **3** to **6** are complex in the fingerprint region and the expected  $\text{U}=\text{O}$  absorption features between 800–700  $\text{cm}^{-1}$  are masked by those of the macrocyclic ligand and the  $\text{O}-\text{SiR}_3$  groups (Supplementary Fig. 1).

We have sought to generalize the reaction further, and have found that the potassium silylamide may be replaced by potassium hydride, another strong base, in combination with other sources of silyl group. Thus, the replacement of  $\text{KN}(\text{Si}(\text{CH}_3)_3)_2$  by KH and either  $\text{N}(\text{Si}(\text{CH}_3)_3)_3$  or  $\text{C}_6\text{H}_5\text{CH}_2\text{Si}(\text{CH}_3)_3$  is equally effective in the synthesis of **3**, affording isolated yields of up to 85%, via N–Si or C–Si bond cleavage (see online Methods). In contrast, however, treatment of **1** with a reductant (rather than a base), and a source of  $\text{Si}(\text{CH}_3)_3$ , in these cases cobaltocene and trimethylsilyl triflate, does not result in reductive silylation.

These data suggest that this new and general reaction to reductively silylate the uranyl oxo group requires the deprotonation of the empty macrocyclic cavity by the potassium base to form potentially an oxidizing,  $\text{U}(\text{VI})$  intermediate **K**<sub>2</sub>-**1** (Fig. 1b) in which the *endo*- $\text{U}=\text{O}$  bond is coordinated by two K cations, and the *exo*- $\text{U}=\text{O}$  bond is now polarized sufficiently to engage in N–Si and C–Si bond cleavage.

Transition metal oxo bonds are weakened when a strong ligand is in the *trans* position (the *trans* influence). In contrast, in uranyl compounds, covalent interactions between the oxo ligands and the metal *f* orbitals mutually strengthen the two *trans*  $\text{U}=\text{O}$  bonds, the inverse *trans* influence<sup>26</sup>. In high-oxidation-state porphyrin-based iron oxo chemistry, tuning the axial ligand markedly alters the reactivity of the electrophilic  $\text{Fe}=\text{O}$  group towards alkane hydroxylation and olefin epoxidation<sup>6</sup>. Likewise, by manipulating the uranyl oxo within the molecular cleft, we have significantly disrupted the overall  $\text{UO}_2$  bonding to activate the *exo* oxo group towards reductive silylation. The ready formation of strong  $\text{O}-\text{Si}$  bonds in **3** to **6** parallels that seen in transition metal oxo chemistry in which hydrogen atom abstraction reactions do not require metal-based radicals, but instead depend on the strength of the bond between the oxidant and the hydrogen atom<sup>9</sup>. Unfortunately, attempts to isolate the proposed **K**<sub>2</sub>-**1** intermediate have been unsuccessful. Thermally stable, pentavalent, functionalized uranyl complexes are most readily isolated by substitution of the two K cations by transition metal halides in a reaction that eliminates KI and forms **3** to **6**. The reaction to afford **3** is equally successful when carried out in the dark, confirming the absence of any photochemically derived reactivity.

We recorded variable temperature magnetic measurements to compare the  $f^1d^6d^6$   $\text{UFe}_2$  system **3** with the  $f^1d^{10}d^{10}$   $\text{UZn}_2$  system



**Figure 2** | X-ray crystal structures of  $[\text{UO}(\text{OSi}(\text{CH}_3)_3)(\text{thf})\text{Fe}_2\text{I}_2(\text{L})]$  and  $[\text{UO}(\text{OSi}(\text{CH}_3)_3)(\text{thf})\text{Zn}_2\text{I}_2(\text{L})]$ . Thermal ellipsoid plot (50% probability displacement) views of (a) **3** and (b) **5**. For clarity, all hydrogen atoms and the minor thf component have been removed.

5. The room-temperature moment of 7.74 BM for **3** (BM = Bohr magnetons), and the Curie–Weiss behaviour (2 to 300 K) suggests the presence of two, high-spin, Fe(II) centres and one  $f^1$  U(v) centre (Supplementary Fig. 2) that are magnetically independent; the thermal variation of the product of molar magnetic susceptibility and temperature,  $\chi_M T$ , is dominated by the magnetic contribution from the Fe ions. In contrast, the magnetic behaviour for **5** (2 to 300 K) should only contain contributions from the U centre<sup>27</sup>; it displays two distinct regions (Supplementary Fig. 2) associated with the depopulation of excited crystal field states of the U(v)  $f^1$  cation and is similar to that observed for the few known organometallic pentavalent uranium complexes<sup>28,29</sup>. The moment at low temperature rises from 0.41 to 1.11 BM and increases to 2.38 BM at high temperature. In contrast, the moment of a U(IV) ( $f^2$ ) system would be expected to be higher at room temperature (3.58 BM), and the reciprocal susceptibility would become temperature-independent below about 40 K. A preliminary electron paramagnetic resonance study of **5** in frozen methyl-thf at 5 K (Supplementary Fig. 3) displays a strong, broad resonance at  $g = 2.2$  that supports the presence of a single  $f$  electron.

We have shown that the use of a macrocyclic architecture to place the uranyl ion in a rigid and asymmetric coordination environment allows the generation of a reactive and highly oxidizing uranyl complex which can selectively cleave N–Si and C–Si bonds to form singly, covalently functionalized pentavalent uranyl complexes. These reactive U oxo compounds may also provide functional chemical models for the highly radioactive  $f^1$  plutonium and neptunium dioxo cations<sup>30</sup>.

## METHODS SUMMARY

Working under a dry, oxygen-free dinitrogen atmosphere, with reagents dissolved or suspended in aprotic solvents, and combined or isolated using cannula and glove box techniques, we first treated the free macrocycle  $H_4L$  with a bis(amido) uranyl precursor, to form the hinged macrocyclic complex  $[UO_2(thf)(H_2L)]$  in which one  $N_4$ -donor compartment remains vacant. Treatment of this complex with two equivalents of potassium base and a suitable silylated reagent (or a base containing an ancillary silyl group) afforded a soluble complex in which the uranium was shown to be both singly reduced and silylated at the *exo* oxo group, as the  $UO(OSiR_3)_2$  dication. This asymmetric pentavalent uranyl complex is then readily isolated, purified, and characterized by a final salt elimination reaction to produce two equivalents of potassium halide, and to place two transition metal cations (as Fe or Zn chloride or iodide salts, MX) into the remaining cavity of the macrocycle, affording  $[UO(OSiR_3)(thf)(L)(MX)_2]$ . We characterized all compounds by elemental analysis, Fourier transform infrared spectroscopy, and either variable-temperature magnetic moment measurements or nuclear magnetic resonance (NMR) spectroscopy (paramagnetic and diamagnetic compounds respectively). Additionally, we determined the solid-state structures of two of the silylated complexes by single-crystal X-ray diffraction studies.

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

Received 8 June; accepted 9 November 2007.

1. Amme, M., Wiss, T., Thiele, H., Boulet, P. & Lang, H. Uranium secondary phase formation during anoxic hydrothermal leaching processes of  $UO_2$  nuclear fuel. *J. Nucl. Mater.* **341**, 209–223 (2005).
2. Lovley, D. R., Phillips, E. J. P., Gorby, Y. A. & Landa, E. R. Microbial reduction of uranium. *Nature* **350**, 413–416 (1991).
3. Suzuki, Y., Kelly, S. D., Kemner, K. M. & Banfield, J. F. Radionuclide contamination: Nanometre-size products of uranium bioreduction. *Nature* **419**, 134 (2002).
4. Denning, R. G. Electronic structure and bonding in actinyl ions and their analogs. *J. Phys. Chem. A* **111**, 4125–4143 (2007).
5. Kühn, F. E., Santos, A. M. & Abrantes, M. Mononuclear organomolybdenum(vi) dioxo complexes: Synthesis, reactivity, and catalytic applications. *Chem. Rev.* **106**, 2455–2475 (2006).
6. Nam, W. High-valent iron(IV)-oxo complexes of heme and non-heme ligands in oxygenation reactions. *Acc. Chem. Res.* **40**, 522–531 (2007).
7. Jin, N., Ibrahim, M., Spiro, T. G. & Groves, J. T. Trans-dioxo manganese(v) porphyrins. *J. Am. Chem. Soc.* **129**, 12416–12417 (2007).
8. Limberg, C. The role of radicals in metal-assisted oxygenation reactions. *Angew. Chem. Int. Edn Engl.* **42**, 5932–5954 (2003).
9. Mayer, J. M. Hydrogen atom abstraction by metal-oxo complexes: Understanding the analogy with organic radical reactions. *Acc. Chem. Res.* **31**, 441–450 (1998).

10. Burdet, F., Pecaut, J. & Mazzanti, M. Isolation of a tetrameric cation-cation complex of pentavalent uranyl. *J. Am. Chem. Soc.* **128**, 16512–16513 (2006).
11. Natrajan, L., Burdet, F., Pecaut, J. & Mazzanti, M. Synthesis and structure of a stable pentavalent-uranyl coordination polymer. *J. Am. Chem. Soc.* **128**, 7152–7153 (2006).
12. Berthet, J. C., Siffredi, G., Thuery, P. & Ephritikhine, M. Easy access to stable pentavalent uranyl complexes. *Chem. Commun.* 3184–3186 (2006).
13. Burns, C. J. *et al.* A trigonal bipyramidal uranyl amido complex: Synthesis and structural characterization of  $Na(thf)_2UO_2\{N(SiMe_3)_2\}_3$ . *Inorg. Chem.* **39**, 5464–5468 (2000).
14. Sarsfield, M. J., Helliwell, M. & Raftery, J. Distorted equatorial coordination environments and weakening of U=O bonds in uranyl complexes containing NCN and NPN ligands. *Inorg. Chem.* **43**, 3170–3179 (2004).
15. Sarsfield, M. J. & Helliwell, M. Extending the chemistry of the uranyl ion: Lewis acid coordination to a U=O oxygen. *J. Am. Chem. Soc.* **126**, 1036–1037 (2004).
16. Kannan, S., Vaughn, A. E., Weis, E. M., Barnes, C. L. & Duval, P. B. Anhydrous photochemical uranyl(vi) reduction: Unprecedented retention of equatorial coordination accompanying reversible axial oxo/alkoxide exchange. *J. Am. Chem. Soc.* **128**, 14024–14025 (2006).
17. Arnold, P. L., Blake, A. J., Wilson, C. & Love, J. B. Uranyl complexation by a Schiff-base, polypyrrrolic macrocycle. *Inorg. Chem.* **43**, 8206–8208 (2004).
18. Arnold, P. L., Patel, D., Blake, A. J., Wilson, C. & Love, J. B. Selective oxo functionalization of the uranyl ion with 3d metal cations. *J. Am. Chem. Soc.* **128**, 9610–9611 (2006).
19. Docrat, T. I. *et al.* X-ray absorption spectroscopy of tricarbonatodioxouranate(v),  $[UO_2(CO_3)_3]^{5-}$ , in aqueous solution. *Inorg. Chem.* **38**, 1879–1882 (1999).
20. Hay, P. J., Martin, R. L. & Schreckenbach, G. Theoretical studies of the properties and solution chemistry of  $AnO_2^{2+}$  and  $AnO^{2+}$  aquo complexes for  $An = U, Np$ , and  $Pu$ . *J. Phys. Chem. A* **104**, 6259–6270 (2000).
21. Wander, M. C. F., Kerisit, S., Rosso, K. M. & Schoonen, M. A. A. Kinetics of tricarbonato uranyl reduction by aqueous ferrous iron: A theoretical study. *J. Phys. Chem. A* **110**, 9691–9701 (2006).
22. Zi, G. *et al.* Preparation and reactions of base-free bis(1,2,4-tri-*tert*-butylcyclopentadienyl)uranium oxide,  $Cp'_2UO$ . *Organometallics* **24**, 4251–4264 (2005).
23. Cotton, F. A., Marler, D. O. & Schwotzer, W. Dinuclear uranium alkoxides: preparation and structures of  $KU_2(OCMe_3)_9$ ,  $U_2(OCMe_3)_9$ , and  $U_2(OCHMe_2)_{10}$ , containing  $[U(IV)U(IV)]$ ,  $[U(IV)U(V)]$ , and  $[U(V)U(V)]$ , respectively. *Inorg. Chem.* **23**, 4211–4215 (1984).
24. Donahue, J. P., Goldsmith, C. R., Nadiminti, U. & Holm, R. H. Synthesis, structures, and reactivity of bis(dithiolene)molybdenum(IV,VI) complexes related to the active sites of molybdoenzymes. *J. Am. Chem. Soc.* **120**, 12869–12881 (1998).
25. Lorber, C., Donahue, J. P., Goddard, C. A., Nordlander, E. & Holm, R. H. Synthesis, structures, and oxo transfer reactivity of bis(dithiolene)tungsten(IV, VI) complexes related to the active sites of tungstoenzymes. *J. Am. Chem. Soc.* **120**, 8102–8112 (1998).
26. O'Grady, E. & Kaltsoyannis, N. On the inverse trans influence. Density functional studies of  $[MOX_5]^{n-}$  ( $M = Pa, n = 2; M = U, n = 1; M = Np, n = 0; X = F, Cl$  or  $Br$ ). *J. Chem. Soc., Dalton Trans.* 1233–1239 (2002).
27. Costes, J. P., Dahan, F., Dupuis, A. & Laurent, J. P. Nature of the magnetic interaction in the  $(Cu^{2+}, Ln^{3+})$  pairs: An empirical approach based on the comparison between homologous  $(Cu^{2+}, Ln^{3+})$  and  $(NiL_5^{2+}, Ln^{3+})$  complexes. *Chem. Eur. J.* **4**, 1616–1620 (1998).
28. Castro-Rodriguez, I., Olsen, K., Gantzel, P. & Meyer, K. Uranium tris-aryloxide derivatives supported by triazacyclononane: engendering U(III) center with a single pocket for reactivity. *J. Am. Chem. Soc.* **125**, 4565–4571 (2003).
29. Rosen, R. K., Andersen, R. A. & Edelstein, N. M. A bimetallic molecule with antiferromagnetic coupling between the uranium centres. *J. Am. Chem. Soc.* **112**, 4588–4590 (1990).
30. Reilly, S. D. & Neu, M. P. Pu(VI) hydrolysis: further evidence for a dimeric plutonyl hydroxide and contrasts with U(VI) chemistry. *Inorg. Chem.* **45**, 1839–1846 (2006).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank the EPSRC (UK), the Royal Society, and the Universities of Edinburgh and Nottingham for support, J. Sanchez-Benitez and P. Anderson of Edinburgh University for help with magnetic susceptibility measurements and chloride analysis respectively, R. Edge and the EPSRC EPR service at the University of Manchester, and D. Leigh for his advice.

**Author Contributions** D.P. synthesized and characterized the compounds, and solved the crystal structure data. C.W. mounted the crystals, collected the single-crystal X-ray crystallographic data, modelled the disorder components in the structures, and checked the final structure solutions. P.L.A. and J.B.L. generated and managed the project, helped characterize the complexes, analysed the data and wrote the manuscript.

**Author Information** X-ray crystallographic coordinates for **3** and **5** have been deposited at the Cambridge Crystallographic Database, numbers 649987 and 649988 respectively. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence and requests for materials should be addressed to P.L.A. (Polly.Arnold@ed.ac.uk) or J.B.L. (Jason.Love@ed.ac.uk).

## METHODS

**[UO<sub>2</sub>(thf)(H<sub>2</sub>L)]·thf, 1.** To a stirred solution of [UO<sub>2</sub>(thf)<sub>2</sub>{N(Si(CH<sub>3</sub>)<sub>3</sub>)<sub>2</sub>}<sub>2</sub>] (2.94 g, 4.0 mmol) in thf (20 ml, −78 °C) we added slowly a solution of H<sub>4</sub>L (2.64 g, 4.0 mmol) in thf (20 ml, −78 °C). The resulting solution was allowed to warm to room temperature over 16 h, after which the volatiles were removed under vacuum and the residual solids redissolved in thf (15 ml). Addition of hexane (20 ml) afforded a precipitate that was isolated by filtration, washed with hexane (2 × 10 ml), and dried under vacuum to yield 3.76 g, 88% of **1** as a brown solid. Analysis. Found: C, 56.00; H, 5.55; N, 10.51. C<sub>50</sub>H<sub>58</sub>N<sub>8</sub>O<sub>4</sub>U requires: C, 55.96; H, 5.46; N, 10.44%; infrared (Nujol, cm<sup>−1</sup>): ν 908(s) (UO<sub>2</sub> asymmetric stretch).

**[UO(OSi(CH<sub>3</sub>)<sub>3</sub>)(thf)Fe<sub>2</sub>I<sub>2</sub>(L)], 3.** To a stirred mixture of **1** (0.27 g, 0.25 mmol) and KN(Si(CH<sub>3</sub>)<sub>3</sub>)<sub>2</sub> (0.10 g, 0.53 mmol) we added thf (20 ml) at −78 °C, and added the resulting solution dropwise to stirred slurry of FeI<sub>2</sub> (0.15 g, 0.50 mmol, beads) in thf (10 ml, −78 °C). The resulting mixture was allowed to warm to room temperature over 42 h, after which we removed the solid KI by filtration and washed it with thf (2 × 5 ml). The combined filtrates were evaporated to dryness, the residual solids extracted into hot toluene (20 ml), filtered and dried under vacuum to yield 0.29 g, 80% of **3** as a dark red solid. Analysis. Found: C, 40.93; H, 4.07; N, 7.64. C<sub>49</sub>H<sub>57</sub>N<sub>8</sub>O<sub>3</sub>Fe<sub>2</sub>I<sub>2</sub>SiU requires: C, 40.93; H, 4.00; N, 7.79%. Magnetic moment (superconducting quantum interference device (SQUID) 300 K): μ<sub>eff</sub> 7.74 BM; electron impact mass spectrometry: *m/z* 343 (37.7%, [UO(OSi(CH<sub>3</sub>)<sub>3</sub>)]<sup>+</sup>).

**Alternative syntheses of 3. A.** To a stirred mixture of **1** (0.10 g, 0.09 mmol) and KH (9 mg, 0.23 mmol) we added thf (20 ml) at −78 °C, and allowed the mixture to warm to room temperature over 45 min. We filtered the resulting mixture dropwise by cannula into a stirred slurry of FeI<sub>2</sub> (56 mg, 0.18 mmol) and N(Si(CH<sub>3</sub>)<sub>3</sub>)<sub>3</sub> (21 mg, 0.09 mmol) in thf (10 ml, −78 °C). Room-temperature work up as above yielded 0.09 g, 69% of **3** as a dark red solid. **B.** To a stirred mixture of **1** (0.10 g, 0.09 mmol) and KH (9 mg, 0.23 mmol) we added thf (20 ml) at −78 °C and allowed the mixture to warm to room temperature over 45 min. We filtered the resulting mixture dropwise on to a stirred slurry of FeI<sub>2</sub> (56 mg, 0.18 mmol) and C<sub>6</sub>H<sub>5</sub>CH<sub>2</sub>Si(CH<sub>3</sub>)<sub>3</sub> (15 mg, 0.09 mmol) in thf (15 ml, −78 °C). Room-temperature work up as above yielded 0.11 g, 85% of **3** as a dark red solid.

**[UO(OSi(CH<sub>3</sub>)<sub>3</sub>)(thf)Zn<sub>2</sub>I<sub>2</sub>(L)], 5.** To a stirred mixture of **1** (0.34 g, 0.32 mmol) and KN(Si(CH<sub>3</sub>)<sub>3</sub>)<sub>2</sub> (0.13 g, 0.63 mmol) we added thf (20 ml) at −78 °C. After 15 min, we added the mixture dropwise to a stirred slurry of ZnI<sub>2</sub> (0.20 g, 0.63 mmol) in toluene (20 ml, −78 °C). Room-temperature work up as above yielded 0.21 g, 46% of **5** as a pale brown solid. Analysis. Found: C, 40.30; H, 3.91;

7.70. C<sub>49</sub>H<sub>57</sub>N<sub>8</sub>I<sub>2</sub>O<sub>3</sub>SiZn<sub>2</sub>U requires: C, 40.40; H, 3.95; N, 7.69%. Magnetic moment (SQUID, 300 K): μ<sub>eff</sub> 2.38 BM. Electron paramagnetic resonance spectroscopy (frozen glass methyl-thf solution, 5 K, 0–1.6 T, 2 mW, 9.610794 GHz): *g* = 2.2.

**[UO(OSi(CH<sub>3</sub>)<sub>3</sub>)(thf)Zn<sub>2</sub>Cl<sub>2</sub>(L)], 6.** To a stirred mixture of **1** (0.10 g, 0.09 mmol) and KN(Si(CH<sub>3</sub>)<sub>3</sub>)<sub>2</sub> (0.036 g, 0.18 mmol) we added thf (15 ml) at −78 °C. After 15 min, we added the mixture dropwise to a stirred slurry of ZnCl<sub>2</sub> (0.025 g, 0.18 mmol) in toluene (20 ml, −78 °C). Room-temperature work up as above yielded 0.06 g, 56% of **6** as a pale brown solid. Analysis. Found: C, 46.30; H, 4.50; 8.72. C<sub>49</sub>H<sub>57</sub>N<sub>8</sub>Cl<sub>2</sub>O<sub>3</sub>SiZn<sub>2</sub>U requires: C, 46.19; H, 4.52; N, 8.80%. Magnetic moment (SQUID, 300 K): μ<sub>eff</sub> 3.01 BM.

**Reaction between 1 and KN(Si(CH<sub>3</sub>)<sub>3</sub>)<sub>2</sub>: attempted synthesis of [UO(OSi(CH<sub>3</sub>)<sub>3</sub>)(thf)K<sub>2</sub>L].** To a stirred mixture of **1** (0.10 g, 0.10 mmol) and KN(Si(CH<sub>3</sub>)<sub>3</sub>)<sub>2</sub> (0.041 g, 0.21 mmol) we added thf (20 ml) at −78 °C. We allowed the resulting red solution to warm to room temperature over 2 h, after which we removed the volatiles from the now dark brown solution. We washed the solid residues with toluene (1 × 10 ml) and dried them to form a dark brown solid, which was redissolved in a minimal amount of thf (1–2 ml) and cooled (−30 °C) for 16 h. The resulting dark precipitate was isolated and was found to be no longer soluble in thf. Elemental analysis indicated that the compound had decomposed.

**Reaction between 1 and cobaltocene and trimethylsilyl triflate: attempted synthesis of [UO(OSi(CH<sub>3</sub>)<sub>3</sub>)(thf)(H<sub>2</sub>L)] and cobaltocenium triflate.** To a stirred mixture of **1** (0.10 g, 0.09 mmol) and Co(C<sub>5</sub>H<sub>5</sub>)<sub>2</sub> (0.017 g, 0.09 mmol) we added thf (20 ml) at −78 °C, and added (CH<sub>3</sub>)<sub>3</sub>SiOTf (0.020 g, 0.09 mmol) into the mixture by syringe. We allowed the mixture to warm to room temperature over 16 h. We removed the volatiles from the now dark red solution to afford a viscous red oil. Elemental analysis indicated that the compounds had decomposed.

**Reaction between 1 and excess KH for the identification of by-products.** We added cold thf (0.5 ml, −35 °C) and a few drops of C<sub>6</sub>D<sub>6</sub> to cold (−35 °C) **1** (10 mg, 0.009 mmol) and KH (2 mg, 0.05 mmol) in a Teflon-tapped NMR tube. Upon warming, we observed gas evolution, which we identified as dissolved dihydrogen at δ = 4.4 p.p.m. in the <sup>1</sup>H NMR spectrum.

**Reaction between 1 and 2 KN(Si(CH<sub>3</sub>)<sub>3</sub>)<sub>2</sub> for the identification of by-products.** We added cold thf (0.5 ml, −35 °C) and a few drops of C<sub>6</sub>D<sub>6</sub> to cold (−35 °C) **1** (5 mg, 0.005 mmol), and KN(Si(CH<sub>3</sub>)<sub>3</sub>)<sub>2</sub> (1.8 mg, 0.009 mmol) in a Teflon-tapped NMR tube. By integration, one molar equivalent of HN(Si(CH<sub>3</sub>)<sub>3</sub>)<sub>2</sub> was observed in the <sup>1</sup>H NMR spectrum.

**Crystallography.** Dark red single crystals of **3** (needle-shaped) and **5** (parallel-piped) were grown from saturated C<sub>6</sub>D<sub>6</sub> solutions at room temperature.

## LETTERS

# Programming biomolecular self-assembly pathways

Peng Yin<sup>1,2</sup>, Harry M. T. Choi<sup>1</sup>, Colby R. Calvert<sup>1</sup> & Niles A. Pierce<sup>1,3</sup>

In nature, self-assembling and disassembling complexes of proteins and nucleic acids bound to a variety of ligands perform intricate and diverse dynamic functions. In contrast, attempts to rationally encode structure and function into synthetic amino acid and nucleic acid sequences have largely focused on engineering molecules that self-assemble into prescribed target structures, rather than on engineering transient system dynamics<sup>1,2</sup>. To design systems that perform dynamic functions without human intervention, it is necessary to encode within the biopolymer sequences the reaction pathways by which self-assembly occurs. Nucleic acids show promise as a design medium for engineering dynamic functions, including catalytic hybridization<sup>3–6</sup>, triggered self-assembly<sup>7</sup> and molecular computation<sup>8,9</sup>. Here, we program diverse molecular self-assembly and disassembly pathways using a ‘reaction graph’ abstraction to specify complementarity relationships between modular domains in a versatile DNA hairpin motif. Molecular programs are executed for a variety of dynamic functions: catalytic formation of branched junctions, autocatalytic duplex formation by a cross-catalytic circuit, nucleated dendritic growth of a binary molecular ‘tree’, and autonomous locomotion of a bipedal walker.

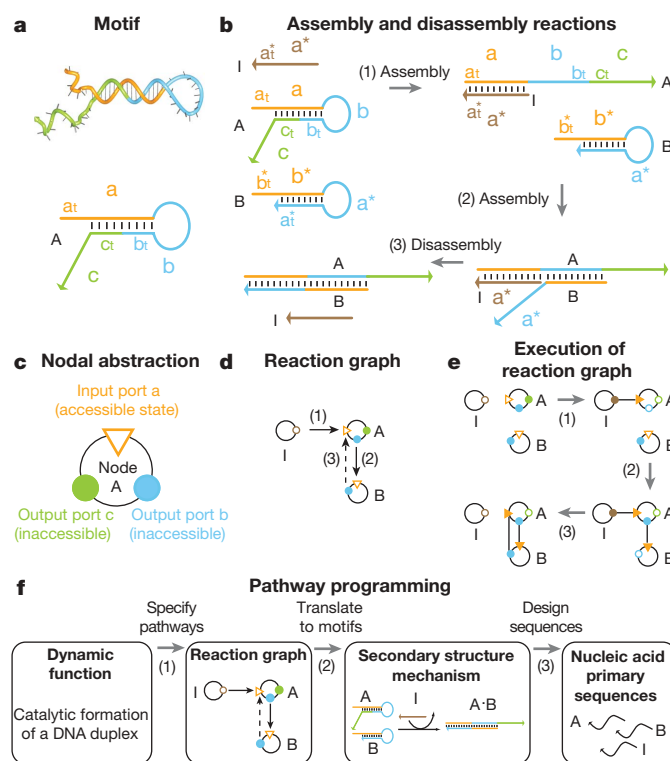
The hairpin motif (A in Fig. 1a) comprises three concatenated domains, a, b and c. Each domain contains a special nucleation site called a toehold<sup>10</sup>, denoted  $a_t$ ,  $b_t$  and  $c_t$ . Two basic reactions can be programmed using this motif, as illustrated for the example of catalytic duplex formation in Fig. 1b. First, an assembly reaction (1) occurs when a single-stranded initiator I, containing an exposed toehold  $a_t^*$ , nucleates at the exposed toehold  $a_t$  of hairpin A, initiating a branch migration that opens the hairpin. Hairpin domains b and c, with newly exposed toeholds  $b_t$  and  $c_t$ , can then serve as assembly initiators for other suitably defined hairpins, permitting cascading (for example, in reaction (2), domain b of hairpin A assembles with domain  $b^*$  of hairpin B, opening the hairpin). Second, a disassembly reaction (3) occurs when a single-stranded domain ( $a^*$  of B) initiates a branch migration that displaces the initiator I from A. In this example, I catalyses the formation of duplex A•B through a prescribed reaction pathway.

To assist in programming more complex reaction pathways, we abstract the motif of Fig. 1a as a node with three ports (Fig. 1c): a triangular input port and two circular output ports. The state of each port is either accessible (open triangle/circle) or inaccessible (solid triangle/circle), depending on whether the toehold of the corresponding motif domain is exposed or sequestered. Functional relationships between ports within a node are implicit in the definition of the nodal abstraction corresponding to a particular motif (for example, for the node of Fig. 1c, the output ports flip to accessible states if the input port is flipped to an inaccessible state through an interaction with a complementary upstream output port). By depicting assembly reactions by solid arrows and disassembly reactions by dashed arrows (each directed from an output port to a complementary input port of a different node), reaction pathways can be

specified abstractly in the form of a reaction graph, representing a program to be executed by nucleic acid molecules.

The reactions depicted in the secondary structure mechanism of Fig. 1b are specified using a reaction graph in Fig. 1d. The initial conditions for this program are described via the state of each port in the reaction graph. Figure 1e depicts the execution of this reaction graph through cascaded assembly and disassembly reactions. An assembly reaction is executed when ports connected by a solid arrow are simultaneously accessible. For the initial conditions depicted in Fig. 1d, the program must start with the execution of reaction (1).

Reaction 1 (assembly): in an assembly reaction (executed here by the accessible output port of I and the complementary accessible input port of A), a bond is made between the ports and they are flipped to inaccessible states; the two output ports of A are flipped



**Figure 1 | Programming biomolecular self-assembly pathways.**

**a**, Secondary structure of the hairpin motif. Coloured lines represent strand domains; short black lines represent base pairs; arrowheads indicate 3' ends. Domain c is optional. **b**, Secondary structure mechanism illustrating assembly and disassembly reactions during catalytic duplex formation. Asterisks denote complementarity. **c**, Abstraction of the motif A as a node with three ports (colour use is consistent with **a**). **d**, A reaction graph representing a molecular program executed schematically in **b** and **e**. **e**, Execution of the reaction graph of **d**. **f**, Hierarchical design process.

<sup>1</sup>Department of Bioengineering, <sup>2</sup>Department of Computer Science, <sup>3</sup>Department of Applied & Computational Mathematics, California Institute of Technology, Pasadena, California 91125, USA.

to accessible states (based on the internal logic of node A). Reaction 2 (assembly): a bond is made between the newly accessible blue output port of A and the complementary accessible input port of B and both ports are flipped to inaccessible states; the output port of B is flipped to the accessible state (based on the internal logic of node B). Reaction 3 (disassembly): in a disassembly reaction (executed here by the newly accessible output port of B, the inaccessible input port of A, and the inaccessible output port of I), the bond between the output port of I and the input port of A is displaced by a bond between the output port of B and the input port of A; the states of the two output ports are flipped (see Supplementary Information 2 for additional details).

The reaction graph provides a simple representation of assembly (and disassembly) pathways that can be translated directly into molecular executables: nodes represent motifs, ports represent domains, states describe accessibility, arrows represent assembly and disassembly reactions between complementary ports. Starting from a conceptual dynamic function, a molecular implementation is realized in three steps (Fig. 1f): (1) pathway specification via a reaction graph; (2) translation into secondary structure motifs; (3) computational design of motif primary sequences (see Methods for details). We demonstrate the utility of this hierarchical design process by experimentally executing molecular programs encoding four distinct dynamic functions.

**Program 1: Catalytic geometry.** Current protocols for self-assembling synthetic DNA nanostructures often rely on annealing procedures to bring interacting DNA strands to equilibrium on the free-energy landscape<sup>11–13</sup>. By contrast, self-assembly in biology proceeds isothermally and assembly kinetics are often controlled by catalysts. Until now, synthetic DNA catalysts<sup>3–6</sup> have been used to control the kinetics of the formation of DNA duplex structures. The next challenge is to catalyse the formation of branched DNA structures, the basic building blocks for DNA structural nanotechnology<sup>14,15</sup>.

First, we demonstrate the catalytic formation of a three-arm DNA junction. The assembly and disassembly pathways specified in the reaction graph of Fig. 2a are translated into the motif-based molecular implementation of Fig. 2b (see Supplementary Information 3.1 for details). The complementarity relationships between the segments of hairpins A, B, and C are specified (Fig. 2b, top) so that in the absence of initiator strand I, the hairpins are kinetically impeded from forming the three-arm junction that is predicted to dominate at equilibrium. In the reaction graph, this property is programmed by the absence of a starting point if node I is removed from the graph (that is, no pair of accessible ports connected by an assembly arrow). The introduction of I into the system (Fig. 2b, bottom) activates a cascade of assembly steps with A, B and C, followed by a disassembly

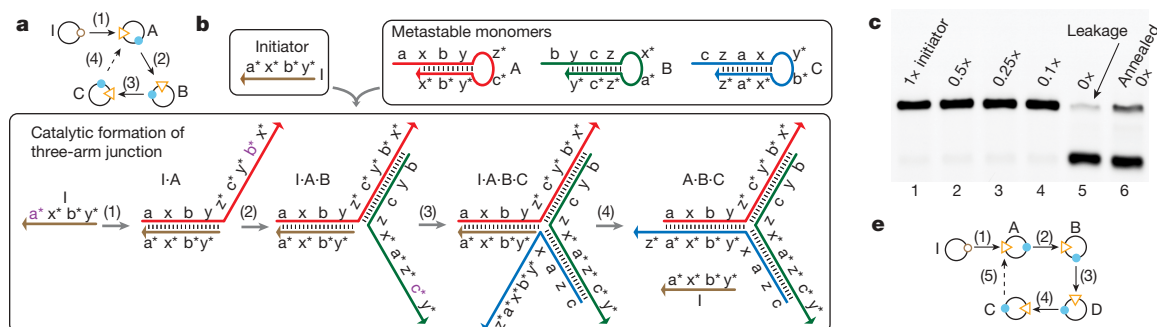
step in which C displaces I from the complex, freeing I to catalyse the self-assembly of additional branched junctions.

Gel electrophoresis confirms that the hairpins assemble slowly in the absence of initiator and that assembly is markedly accelerated by the addition of initiator (Fig. 2c). Disassembly of the initiator leads to catalytic turnover, as indicated by the nearly complete consumption of hairpins even at substoichiometric initiator concentrations. Interestingly, only minimal assembly is achieved by annealing the hairpin mixture, illustrating the utility of pathway programming for traversing free-energy landscapes with kinetic traps that cannot be overcome by traditional annealing approaches.

Direct imaging of the catalysed self-assembly product  $A \cdot B \cdot C$  by atomic force microscopy (AFM) reveals the expected three-arm junction morphology (Fig. 2d). In principle, the reaction pathway can be extended to the catalytic self-assembly of  $k$ -arm junctions (Supplementary Information 3.5). We illustrate  $k = 4$  with the reaction graph and AFM image of Fig. 2e and f.

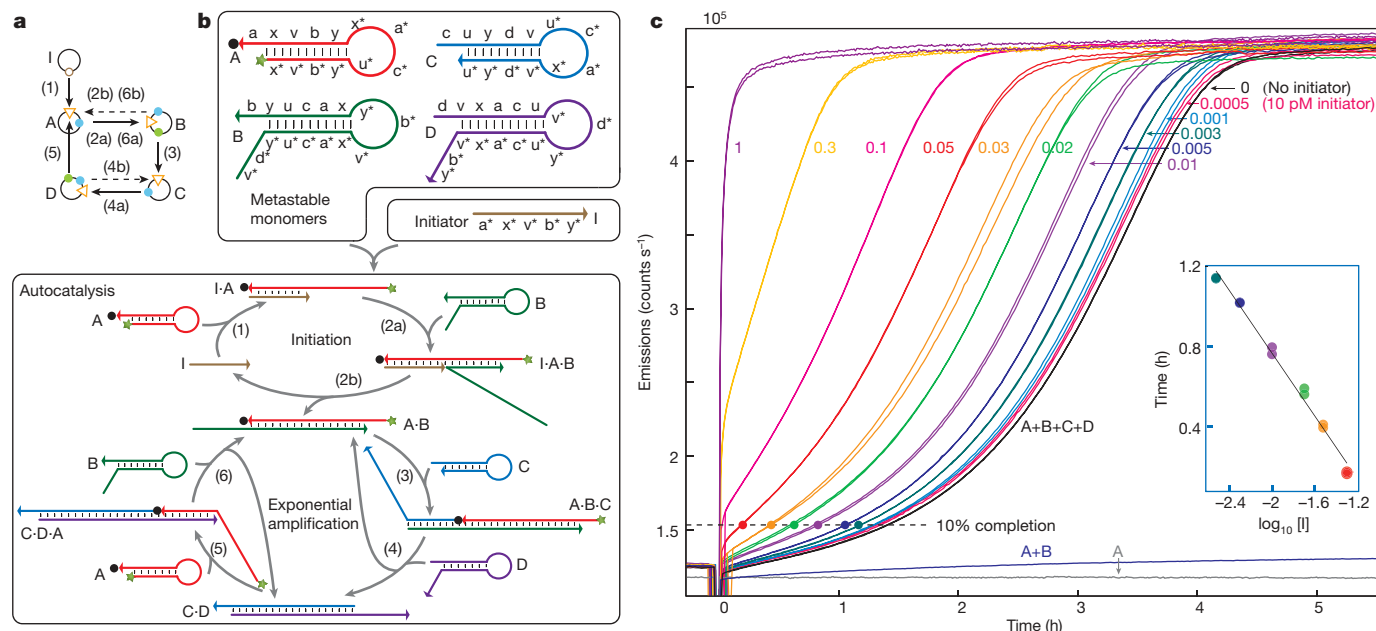
**Program 2: Catalytic circuitry.** By programming cross-catalytic self-assembly pathways in the reaction graph of Fig. 3a, we obtain an autocatalytic system with exponential kinetics. In the corresponding molecular implementation, four hairpin species, A, B, C and D, coexist metastably in the absence of initiator I (Fig. 3b, top). The initiator catalyses the assembly of hairpins A and B to form duplex  $A \cdot B$  (steps 1–2, Fig. 3b, bottom), bringing the system to an exponential amplification stage powered by a cross-catalytic circuit: the duplex  $A \cdot B$  has a single-stranded region that catalyses the assembly of C and D to form  $C \cdot D$  (steps 3–4); duplex  $C \cdot D$  in turn has a single-stranded region that is identical to I and can thus catalyse A and B to form  $A \cdot B$  (steps 5–6). Hence,  $A \cdot B$  and  $C \cdot D$  form an autocatalytic set capable of catalysing its own production. Disassembly (steps 2b, 4b and 6b) is fundamental to the implementation of autocatalysis and sterically uninhibited exponential growth.

Each step in the reaction is examined using native polyacrylamide gel electrophoresis (Supplementary Fig. 12), showing the expected assembly and disassembly behaviour. System kinetics are examined in a fluorescence quenching experiment (Fig. 3c). Spontaneous initiation in the absence of initiator reflects the finite timescale associated with the metastability of the hairpins and yields a sigmoidal time course characteristic of an autocatalytic system<sup>16</sup>. As expected, the curve shifts to the left as the concentration of initiator is increased. A plot of 10% completion time against the logarithm of the concentration shows a linear regime, consistent with exponential kinetics and analytical modelling (Fig. 3c, inset). The minimal leakage of a system containing only A and B (labelled A + B in Fig. 3c) emphasizes that the sigmoidal kinetics of spontaneous initiation for the full system ( $A + B + C + D$ ) are due to cross-catalysis.



**Figure 2 | Programming catalytic geometry: catalytic self-assembly of three-arm and four-arm branched junctions.** See Supplementary Information 3 for details. **a**, Reaction graph for three-arm junctions. **b**, Secondary structure mechanism. Each letter-labelled segment is six nucleotides in length. The initially accessible ( $a^*$  for step 1) or newly exposed ( $b^*$  for Step 2,  $c^*$  for step 3) toeholds that mediate assembly reactions are labelled with purple letters. **c**, Agarose gel electrophoresis demonstrating

catalytic self-assembly for the three-arm system with 750-nM hairpins. Nearly complete conversion of hairpins to reaction products using stoichiometric or substoichiometric initiator I (lanes 1–4). Minimal conversion in the absence of initiator (lane 5), even with annealing (lane 6). **d**, AFM image of a three-arm junction. Scale bar: 10 nm. **e**, Reaction graph and **f**, AFM image for a four-arm junction. Scale bar: 10 nm.



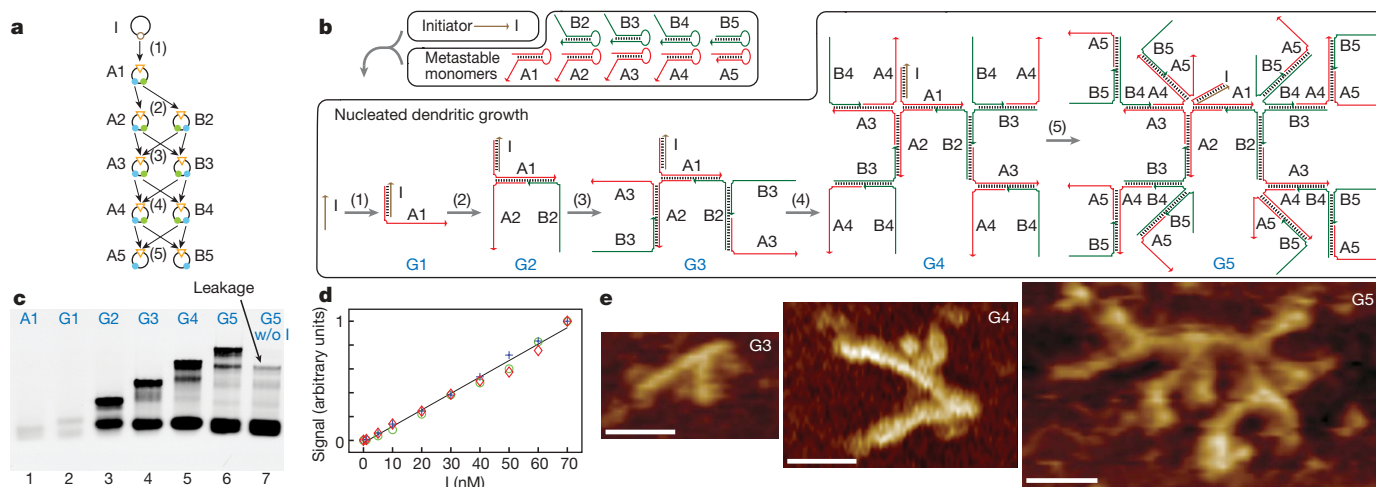
**Figure 3 | Programming catalytic circuitry: autocatalytic duplex formation by a cross-catalytic circuit with exponential kinetics.** See Supplementary Information 4 for details. **a**, Reaction graph. Multiple assembly arrows entering the same input port depict parallel processes on separate copies of the nodal species. **b**, Secondary structure mechanism. **c**, System kinetics examined by fluorescence quenching. Formation of  $A \cdot B$  is monitored by the increase in fluorescence resulting from increased spatial separation between the fluorophore (green star in **b**) and the quencher (black dot in **b**) at either

end of  $A$ . Raw data for two independent reactions are displayed for each initiator concentration (20-nM hairpins). Single traces are shown for the controls containing only  $A$  and  $B$  or only  $A$ . Inset: linear fit of the 10% completion time against the logarithm of the relative concentration of  $I$  ( $0.003 \times \leq [I] \leq 0.05 \times$ ). High-concentration end points ( $[I] \geq 0.1 \times$ ) are excluded based on theoretical analysis; low-concentration end points ( $[I] \leq 0.001 \times$ ) are excluded because of signal poisoning by leakage. See Supplementary Information 4.4 for a detailed treatment.

This system demonstrates synthetic biomolecular autocatalysis<sup>17–20</sup> driven by the free energy of base-pair formation. Autocatalysis and exponential system kinetics can also be achieved through entropy-driven hybridization mechanisms<sup>21</sup>. For sensing applications, the triggered exponential growth of these systems suggest the possibility of engineering enzyme-free isothermal detection methods.

Program 3: Nucleated dendritic growth. The molecular program in Fig. 4a depicts the triggered self-assembly of a binary molecular tree of a prescribed size. The reaction starts with the assembly of an

initiator node  $I$  with a root node  $A1$ . Each assembled node subsequently assembles with two child nodes during the next generation of growth, requiring two new node species per generation. In the absence of steric effects, a  $G$ -generation dendrimer requires  $2G-1$  node species and yields a binary tree containing  $2^{G-1}$  monomers, that is, a linear increase in the number of node species yields an exponential increase in the size of the dendrimer product. Figure 4b depicts the motif based implementation of the program depicted in Fig. 4a: hairpins are metastable in the absence of initiator; the



**Figure 4 | Programming nucleated dendritic growth: triggered assembly of quantized binary molecular trees.** See Supplementary Information 5 for details. **a**, Reaction graph. Multiple assembly arrows entering the same input port depict parallel processes on separate copies of the nodal species. **b**, Secondary structure mechanism. **c**, Agarose gel electrophoresis demonstrating triggered self-assembly. Lanes 1–6: the dominant reaction band shifts with the addition of each generation of hairpins. Subdominant

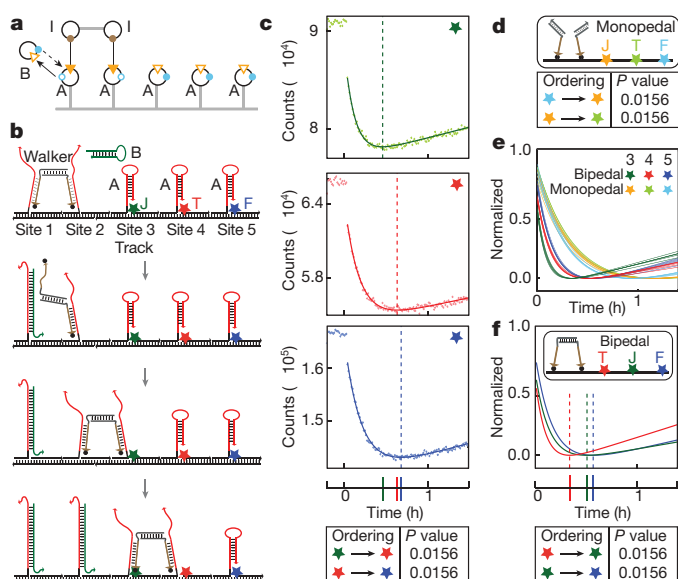
bands are presumed to represent imperfect dendrimers. Lane 7: minimal conversion to reaction products in the absence of initiator. Hairpins  $A1$ ,  $A2$ ,  $B2$  at 62.5 nM; the concentration doubles for each subsequent generation of hairpins. Initiator  $I$  at 50 nM. **d**, Linear relationship between amplification signal (putative  $G5$  reaction product) and initiator for three independent experiments (cross, diamond, circle). See Supplementary Fig. 17 for details. **e**, AFM images of  $G3$ ,  $G4$  and  $G5$  dendrimers. Scale bars: 30 nm.

initiator I triggers the growth of a dendrimer with five generations of branching (G5).

We constructed trees with  $G = 1, 2, 3, 4$  and  $5$ . The nucleated growth of the tree is examined using native agarose gel electrophoresis. Band shifting demonstrates increasing dendrimer size with each generation of growth (Fig. 4c). Figure 4d demonstrates that the concentration of dendrimer depends linearly on the concentration of the initiator in the system. Finally, AFM imaging of dendrimers for  $G = 3, 4$  and  $5$  reveals the expected morphologies (Fig. 4e). Measurements of the dendrimer segment lengths agree well with the design (Supplementary Information 5.4).

In contrast to previous work in which DNA dendrimer target structures were synthesized by sequential ligation of structural subunits<sup>22</sup>, here we program self-assembly pathways so that DNA monomers form dendrimers only on detection of a target nucleation molecule. By growing to a prescribed size, these dendrimers provide quantitative signal amplification with strength exponential in the number of constituent species.

**Program 4: Autonomous locomotion.** The challenge of engineering molecular machines capable of nanoscale autonomous locomotion has attracted much interest in recent years<sup>23–27</sup>. Inspired by



**Figure 5 | Programming autonomous locomotion: stochastic movement of a bipedal walker.** See Supplementary Information 6 for details. **a**, Reaction graph. Bonds between output ports on I and input ports on A represent initial conditions. Static structural elements are depicted by grey line segments. **b**, Secondary structure mechanism depicting processive locomotion. See Supplementary Information 6.1 and 6.3 for non-processive trajectories. **c–f**, Fluorescence quenching experiments measuring the proximity of the quenchers (black dots) on the walker feet to the fluorophores (coloured stars) decorating the track. Fitted curves (solid) are used to determine the time at which the minimum fluorescence (maximum quenching) was observed (dashed vertical line) for each fluorophore. **c**, Bipedal walker with track labelled by fluorophores JOE (green star) → TAMRA (red) → FAM (blue) as in **b**. For each pair of consecutive minima (JOE → TAMRA and TAMRA → FAM), we test the null hypothesis that the median time difference between the minima is zero against the alternative hypothesis that the time difference is positive. Based on a statistical analysis of six independent experiments (see Supplementary Information 6.6, 6.7), the null hypothesis can be rejected for both time differences with the same  $P$ -value of 0.0156, supporting the interpretation that the observed minima are sampled from a distribution in which the ordering of the minima matches the physical ordering of the fluorophores along the track. Similar interpretations apply to the ordering of minima for **d** and **f**. **d**, Monopodal walkers on the same track (JOE (orange star) → TAMRA (pale green) → FAM (pale blue)). **e**, Comparison of time scales for bipedal and monopodal walkers (eighteen traces per walker type: three fluorophores, six experiments). **f**, Bipedal walker with track labelled TAMRA (red star) → JOE (green) → FAM (blue).

the bipedal motor protein, kinesin, which hauls intracellular cargo by striding along microtubules<sup>28</sup>, we have developed an autonomous enzyme-free bipedal DNA walker capable of stochastic locomotion along a DNA track.

Joined by a duplex torso, each of two identical walker legs, I, is capable of catalysing the formation of waste duplex  $A \cdot B$  from metastable fuel hairpins A and B through a reaction pathway in which I assembles with A, which assembles with B, which subsequently disassembles I from the complex (see Fig. 5a and b for the reaction graph and corresponding molecular implementation). The track consists of five A hairpins arranged linearly at regular intervals along a nicked DNA duplex. In the presence of hairpin B, a subpopulation of walkers is expected to move unidirectionally along the track by sequentially catalysing the formation of  $A \cdot B$ . Because of the one-dimensional arrangement of anchor sites, this processive motion occurs only for those walkers that use a foot-over-foot gait by stochastically lifting the back foot at each step.

We investigate walker locomotion using a bulk fluorescence assay that tests whether there is a subpopulation of walkers that moves processively through positions 3, 4 and 5, starting from an initial condition with legs anchored at positions 1 and 2. Quenchers are attached to the walker's legs and spectrally distinct fluorophores are positioned proximal to anchorages 3, 4 and 5. Consistent with processivity, the anticipated sequential transient quenching of the fluorophores at positions 3, 4 and 5 is observed (Fig. 5c). To rule out the possibility that this signal arises from non-processive walker diffusion through the bulk solution from one position to the next, we repeated the experiments using monopodal walkers that lack a mechanism for achieving processivity. In this case, the sequential transient quenching no longer matches the ordering of the fluorophores along the track (Fig. 5d) and the timescale for visiting any one of the three anchorages is longer than the timescale to visit all three anchorages for the bipedal system (Fig. 5e). Additional control experiments (Supplementary Information 6.9) show that this difference in time-scales cannot be explained by the relative rates with which freely diffusing bipedal and monopodal walkers land on the track. As a further test of processivity for the bipedal walker, reordering the fluorophores along the track leads to the expected change in the ordering of the transient quenching (Fig. 5f).

The experimental execution of these four molecular programs demonstrates that the hairpin motif functions as a modular programmable kinetic trap, and that rewiring the connections between nodes in the reaction graph corresponds to rewiring the connections between kinetic traps in the underlying free-energy landscape. In the physical systems, metastable hairpins are initially caught in engineered kinetic traps; the introduction of initiator molecules begins a chain reaction of kinetic escapes in which the hairpin species interact through programmed assembly and disassembly steps to implement dynamic functions. It is important that the timescale of metastability for kinetically trapped molecules is longer than the timescale relevant for the execution of the program. We found it helpful to incorporate clamping segments at the ends of helices to discourage the initiation of non-toehold-mediated branch migrations (see Supplementary Information 3.1). We also found that impure strand syntheses artificially reduce the strength of metastable traps and increase leakage rates. System fidelity was improved by ligating hairpins out of two shorter segments to increase strand purity (Supplementary Information 7.1).

Reaction graphs can be extended beyond the present versatile motif by defining new nodal species that abstract the functional relationships between domains in other motifs. The present hierarchical approach to encoding dynamic function in nucleic acid sequences represents a promising step towards the goal of constructing a compiler for biomolecular function—an automated design process that requires as input a modular conceptual system design, and provides as output a set of biopolymer sequences that encode the

desired dynamic system behaviour (Supplementary Information 7.2).

## METHODS SUMMARY

Starting from a conceptual dynamic function, a molecular implementation is realized in three steps summarized in Fig. 1f. See Supplementary Information 3.1 for an example illustrating the design of the catalytic three-arm junction system. Step (1): pathway specification. We specify the pathway that implements a target dynamic function using a reaction graph. Step (2): translation to motifs. The reaction graph is directly translated to motif secondary structures. First, the basic complementarity requirements are defined and then clamping/padding segments are added (as in Supplementary Information 3.1). Initial dimensioning of the number of nucleotides in each segment is performed using the NUPACK server ([www.nupack.org](http://www.nupack.org)), which models the behaviour of strand species in the context of a dilute solution (including unintended species of complexes)<sup>29</sup>. Step (3): sequence design. Sequences are designed by considering a suite of structures that punctuate the intended reaction pathway or that explicitly preclude undesired off-pathway interactions (for example, structures specifying the absence of an interaction between two strands that should not pair). The sequences are optimized computationally (J. N. Zadeh and R. M. Dirks, personal communication) to maximize affinity and specificity for this suite of structures by minimizing the average number of incorrectly paired bases at equilibrium<sup>30</sup>. We then synthesize and verify the system using gel electrophoresis, bulk fluorescence quenching, or single-molecule AFM.

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

Received 20 July; accepted 31 October 2007.

- Butterfoss, G. L. & Kuhlman, B. Computer-based design of novel protein structures. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 49–65 (2006).
- Seeman, N. C. DNA in a material world. *Nature* **421**, 427–431 (2003).
- Turberfield, A. J. *et al.* DNA fuel for free-running nanomachines. *Phys. Rev. Lett.* **90**, 118102 (2003).
- Bois, J. S. *et al.* Topological constraints in nucleic acid hybridization kinetics. *Nucleic Acids Res.* **33**, 4090–4095 (2005).
- Green, S. J., Lubrich, D. & Turberfield, A. J. DNA hairpins: Fuel for autonomous DNA devices. *Biophys. J.* **91**, 2966–2975 (2006).
- Seelig, G., Yurke, B. & Winfree, E. Catalyzed relaxation of a metastable DNA fuel. *J. Am. Chem. Soc.* **128**, 12211–12220 (2006).
- Dirks, R. M. & Pierce, N. A. Triggered amplification by hybridization chain reaction. *Proc. Natl Acad. Sci. USA* **101**, 15275–15278 (2004).
- Rothmund, P. W. K., Papadakis, N. & Winfree, E. Algorithmic self-assembly of DNA Sierpinski triangles. *PLoS Biol.* **2**, 2041–2053 (2004).
- Seelig, G., Soloveichik, D., Zhang, D. Y. & Winfree, E. Enzyme-free nucleic acid logic circuits. *Science* **314**, 1585–1588 (2006).
- Yurke, B., Turberfield, A. J., Mills, J. A. P., Simmel, F. C. & Neumann, J. L. A DNA-fuelled molecular machine made of DNA. *Nature* **406**, 605–608 (2000).
- Winfree, E., Liu, F., Wenzler, L. A. & Seeman, N. C. Design and self-assembly of two-dimensional DNA crystals. *Nature* **394**, 539–544 (1998).
- Shih, W. M., Quispe, J. D. & Joyce, G. F. A 1.7-kilobase single-stranded DNA that folds into a nanoscale octahedron. *Nature* **427**, 618–621 (2004).
- Rothmund, P. W. K. Folding DNA to create nanoscale shapes and patterns. *Nature* **440**, 297–302 (2006).
- Seeman, N. C. Nucleic acid junctions and lattices. *J. Theor. Biol.* **99**, 237–247 (1982).
- Feldkamp, U. & Niemeyer, C. M. Rational design of DNA nanoarchitectures. *Angew. Chem. Int. Edn Engl.* **45**, 1856–1876 (2006).
- Robertson, A., Sinclair, A. J. & Philp, D. Minimal self-replicating systems. *Chem. Soc. Rev.* **29**, 141–152 (2000).
- von Kiedrowski, G. A self-replicating hexadeoxynucleotide. *Angew. Chem. Int. Edn Engl.* **25**, 932–935 (1986).
- Paul, N. & Joyce, G. F. A self-replicating ligase ribozyme. *Proc. Natl Acad. Sci. USA* **99**, 12733–12740 (2002).
- Levy, M. & Ellington, A. D. Exponential growth by cross-catalytic cleavage of deoxyribozymes. *Proc. Natl Acad. Sci. USA* **100**, 6416–6421 (2003).
- Lee, D. H., Granja, J. R., Martinez, J. A., Severin, K. & Ghadiri, M. R. A self-replicating peptide. *Nature* **382**, 525–528 (1996).
- Zhang, D. Y., Turberfield, A. J., Yurke, B. & Winfree, E. Engineering entropy-driven reactions and networks catalyzed by DNA. *Science* **318**, 1121–1125 (2007).
- Li, Y. *et al.* Controlled assembly of dendrimer-like DNA. *Nature Mater.* **3**, 38–42 (2004).
- Yin, P., Yan, H., Daniell, X. G., Turberfield, A. J. & Reif, J. H. A unidirectional DNA walker that moves autonomously along a track. *Angew. Chem. Int. Edn Engl.* **43**, 4906–4911 (2004).
- Tian, Y., He, Y., Chen, Y., Yin, P. & Mao, C. A DNzyme that walks processively and autonomously along a one-dimensional track. *Angew. Chem. Int. Edn Engl.* **44**, 4355–4358 (2005).
- Bath, J., Green, S. J. & Turberfield, A. J. A free-running DNA motor powered by a nicking enzyme. *Angew. Chem. Int. Edn Engl.* **44**, 4358–4361 (2005).
- Pei, R. *et al.* Behavior of polycatalytic assemblies in a substrate-displaying matrix. *J. Am. Chem. Soc.* **128**, 12693–12699 (2006).
- Venkataraman, S., Dirks, R. M., Rothmund, P. W. K., Winfree, E. & Pierce, N. A. An autonomous polymerization motor powered by DNA hybridization. *Nature Nanotechnol.* **2**, 490–494 (2007).
- Asbury, C. L. Kinesin: world's tiniest biper. *Curr. Opin. Cell Biol.* **17**, 89–97 (2005).
- Dirks, R. M., Bois, J. S., Schaeffer, J. M., Winfree, E. & Pierce, N. A. Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.* **49**, 65–88 (2007).
- Dirks, R. M., Lin, M., Winfree, E. & Pierce, N. A. Paradigms for computational nucleic acid design. *Nucleic Acids Res.* **32**, 1392–1403 (2004).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank the following for discussions: J. S. Bois, R. M. Dirks, M. Grazier G'Sell, R. F. Hariadi, J. A. Othmer, J. E. Padilla, P. W. K. Rothmund, T. Schneider, R. Schulman, M. Schwarzkopf, G. Seelig, D. Sprinzak, S. Venkataraman, E. Winfree, J. N. Zadeh and D. Y. Zhang. We also thank J. N. Zadeh, R. M. Dirks and J. M. Schaeffer for the use of unpublished software, and R. F. Hariadi and S. H. Park for advice on AFM imaging. This work is funded by the NIH, the NSF, the Caltech Center for Biological Circuit Design, the Beckman Institute at Caltech, and the Gates Grubstake Fund at Caltech.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details accompany the paper on Nature's website (<http://www.nature.com/nature>). Correspondence and requests for materials should be addressed to N.A.P. (niles@caltech.edu).

## METHODS

**System design.** A molecular implementation is realized in three steps summarized in Fig. 1f and illustrated in Supplementary Information 3.1. Step (1): pathway specification. Step (2): translation to motifs. Following initial dimensioning using the NUPACK server, the segment dimensions are sometimes further optimized based on subsequent experimental testing. Step (3): sequence design. After computational optimization, occasional further manual optimization was performed using the same design metric on a subset of crucial target structures. We then further analysed the thermodynamic behaviour of the sequences using the NUPACK server. For some systems, stochastic kinetic simulations<sup>31</sup> (J. M. Schaeffer, personal communication) were carried out to confirm the absence of significant kinetic traps along the target reaction pathways. The sequences are shown in Supplementary Information 8.

**System synthesis.** DNA was synthesized and purified by Integrated DNA Technologies. The purified DNA strands were reconstituted in ultrapure water (resistance of 18 M $\Omega$  cm). We determined the concentrations of the DNA solutions by measuring ultraviolet light absorption at 260 nm.

Hairpins were synthesized as two pieces which were then ligated to produce the full hairpin (see Supplementary Information 7.1 for details). We performed the ligation using T4 DNA ligase (New England Biolabs) at either room temperature or 16 °C for a minimum of 2 h. We further purified ligated strands using denaturing polyacrylamide gel electrophoresis. The bands corresponding to the DNA strands of expected sizes were visualized by ultraviolet shadowing and excised from the gel. The DNA strands were then eluted and recovered by ethanol precipitation.

For monomer preparation, we diluted the concentrated DNA strands to reaction conditions: 50 mM Na<sub>2</sub>HPO<sub>4</sub>, 0.5 M NaCl, pH = 6.8 for species in Fig. 2 and Supplementary Fig. 4; and 20 mM Tris, pH = 7.6, 2 mM EDTA, 12.5 mM Mg<sup>2+</sup> (1 × TAE/Mg<sup>2+</sup> buffer) for species in Fig. 3, Supplementary Fig. 12, and Fig. 4. We then annealed the hairpins by heating for 5 min at 90 °C, and then turning off the heating block to allow the system to cool to room temperature (requiring at least 2 h). For walker system assembly, see Supplementary Information 6.4.

**Gel electrophoresis.** For the gel in Fig. 2c, 12  $\mu$ l of each 3- $\mu$ M hairpin species were mixed by pipetting. Portions of this master mix were aliquoted into five separate tubes (6  $\mu$ l per tube). To these tubes we added 2  $\mu$ l of either 3  $\mu$ M I (lane 1), 1.5  $\mu$ M I (lane 2), 0.75  $\mu$ M I (lane 3), 0.3  $\mu$ M I (lane 4), or 1 × reaction buffer (50 mM Na<sub>2</sub>HPO<sub>4</sub>, 0.5 M NaCl, pH = 6.8) (lane 5) to reach a total reaction volume of 8  $\mu$ l. The samples were then mixed by pipetting and allowed to react for 2.5 h at room temperature. The annealed reaction (lane 6), prepared 0.5 h in advance, was made by mixing 2  $\mu$ l of each hairpin with 2  $\mu$ l of the 1 × reaction buffer, and then annealing as described in monomer preparation. A 2% native agarose gel was prepared for use in 1 × LB buffer (Faster Better Media, LLC). We then mixed 1  $\mu$ l of each sample with 1  $\mu$ l of 5 × SYBR Gold loading buffer: 50% glycerol/50% H<sub>2</sub>O/SYBR Gold (Invitrogen) and loaded this into the gel. The gel was run at 350 V for 10 min at room temperature and imaged using an FLA-5100 imaging system (Fuji Photo Film).

For the gel in Fig. 4c, we annealed the hairpins at the following concentrations: A1, A2, B2, A3 and B3 at 1  $\mu$ M; A4 and B4 at 2  $\mu$ M; A5 and B5 at 4  $\mu$ M. The initiator I was prepared at 800 nM. The following sample mixtures were prepared: lane 1, A1; lane 2, I + A1; lane 3, I + A1 + A2 + B2; lane 4, I + A1 + A2 + B2 + A3 + B3; lane 5, I + A1 + A2 + B2 + A3 + B3 + A4 + B4; lane 6, I + A1 + A2 + B2 + A3 + B3 + A4 + B4 + A5 + B5; lane 7, A1 + A2 + B2 + A3 + B3 + A4 + B4 + A5 + B5. Here, I, A1, A2 and B2 were added at 1  $\mu$ l; A3, B3, A4, B4, A5 and B5 at 2  $\mu$ l. We added 1 × reaction buffer (20 mM Tris, pH = 7.6, 2 mM EDTA, 12.5 mM Mg<sup>2+</sup>) to bring the total volume of each sample to 16  $\mu$ l. We mixed the samples by pipetting and allowed them to react for 2 h at room temperature. A 1% native agarose gel was prepared in 1 × LB buffer. We added 8  $\mu$ l of each sample to 2  $\mu$ l 5 × SYBR Gold loading buffer and loaded 8  $\mu$ l of this sample/loading-buffer mix into the gel. The gel was run at 350 V for 10 min at room temperature and then imaged using an FLA-5100 imaging system. For the reactions in Fig. 4d, the hairpins were mixed to reach the following final concentration: A1–C5 (see Supplementary Information 8.4),

A2, B2, 100 nM; A3, B3, 200 nM; A4, B4, 400 nM; A5, B5, 800 nM. We then aliquoted portions of this mix into 10 separate tubes (9  $\mu$ l per tube). To these tubes we added either 1 × TAE/Mg<sup>2+</sup> reaction buffer or the initiator I to give the indicated final concentration of I and a final volume of 11  $\mu$ l. The samples were mixed by pipetting and allowed to react for 1 h at room temperature. We then mixed the sample with 5 × LB loading buffer (Faster Better Media, LLC) to reach 1 × loading buffer concentration (8  $\mu$ l sample, 2  $\mu$ l loading buffer). We loaded the sample/loading buffer mix into a 1% native agarose gel prepared in 1 × LB buffer. The gel was run at 350 V for 10 min at room temperature and then imaged and quantified using an FLA-5100 imaging system. The experiments were performed with 10  $\mu$ M inert 25-nt poly-T carrier strands<sup>21</sup> in the reaction solution. **AFM imaging.** We obtained AFM images using a multimode scanning probe microscope (Veeco Instruments), equipped with a Q-Control module for analogue AFM systems (Atomic Force F&E). The images were obtained in liquid phase under tapping mode using DNP-S oxide sharpened silicon nitride cantilevers (Veeco). We first diluted samples in 1 × TAE/Mg<sup>2+</sup> buffer to achieve the desired imaging density. We applied a 20  $\mu$ l drop of 1 × TAE/Mg<sup>2+</sup> and a 5  $\mu$ l drop of sample to the surface of freshly cleaved mica and allowed them to bind for approximately 2 min. We added supplemental Ni<sup>2+</sup> (15–30 mM) to increase the strength of DNA–mica binding<sup>32</sup>. Before placing the fluid cell on top of the mica puck, we added an additional 15–20  $\mu$ l of 1 × TAE/Mg<sup>2+</sup> buffer to the cavity between the fluid cell and the AFM cantilever chip to avoid bubbles.

**Fluorescence experiments.** For catalytic circuitry experiments, we obtained fluorescence data using a QM-6/2005 steady state spectrofluorometer (Photon Technology International), equipped with a Turret 400™ four-position cuvette holder (Quantum Northwest) and 3.5-ml QS quartz cuvettes (Hellma). The temperature was set to 25 °C. We set the excitation and emission wavelengths to 520 nm (2-nm bandwidth) and 540 nm (4-nm bandwidth), respectively. For the experiments in Fig. 3c, we prepared hairpin monomers, A, B, C and D, and initiator, I, separately as described above. We added 40  $\mu$ l 1- $\mu$ M A to 1.8 ml 1 × TAE/Mg<sup>2+</sup> buffer and mixed it by rapid pipetting eight times using a 1-ml tip. We recorded the baseline signal for ~16 min. Then we added 40  $\mu$ l of 1- $\mu$ M B, C and D and the appropriate concentration of I (or 1 × TAE/Mg<sup>2+</sup> buffer in the case of 0 × I) to the cuvette (to reach the target concentrations described in Fig. 3c) and mixed by rapid pipetting eight times using a 1-ml tip. The control with 20-nM A alone was monitored continuously. The final volume was 2 ml for all experiments. We carried out the experiments with 10- $\mu$ M inert 25-nt poly-T carrier strand<sup>21</sup> in the individual hairpin and initiator stock solutions and ~1- $\mu$ M inert 25-nt poly-T carrier strands in the final reaction solution.

For autonomous locomotion experiments, we used the same spectrofluorometer as above with the temperature controller set to 21 °C. We used two 3.5-ml QS quartz cuvettes (Hellma) in each set of experiments. Excitation and emission wavelengths were set to 492 nm and 517 nm (for FAM), 527 nm and 551 nm (for JOE), and 558 nm and 578 nm (for TAMRA), respectively, with 4-nm bandwidths. The assembly of the walker system is described in Supplementary Information 6.4. We snap-cooled hairpin B in the reaction buffer (4 mM MgCl<sub>2</sub>, 15 mM KCl and 10 mM Tris-HCl, pH = 8.0): heating at 95 °C for 90 s, rapid cooling at room temperature, sitting at room temperature for 30 min before use. The system was assembled using 4 nM track and 3.5 nM bipedal walker. We used a substoichiometric amount of walker to ensure that no free-floating walker would bind to hairpin A on the track. For the same reason, we used substoichiometric monopedal walker (7 nM) in the diffusion experiments. The final concentration of hairpin B was 20 nM, which was equimolar with the five A hairpins on the track (5 × 4 nM = 20 nM). The assembled track was first introduced to record the fluorescence baselines for FAM, JOE and TAMRA. We then introduced hairpin B and mixed 100 times by rapid pipetting to start walker locomotion.

31. Flamm, C., Fontana, W., Hofacker, I. L. & Schuster, P. RNA folding at elementary step resolution. *RNA* **6**, 325–338 (2000).

32. Hansma, H. G. & Laney, D. E. DNA binding to mica correlates with cationic radius: assay by atomic force microscopy. *Biophys. J.* **70**, 1933–1939 (1996).

# Net production of oxygen in the subtropical ocean

Stephen C. Riser<sup>1</sup> & Kenneth S. Johnson<sup>2</sup>

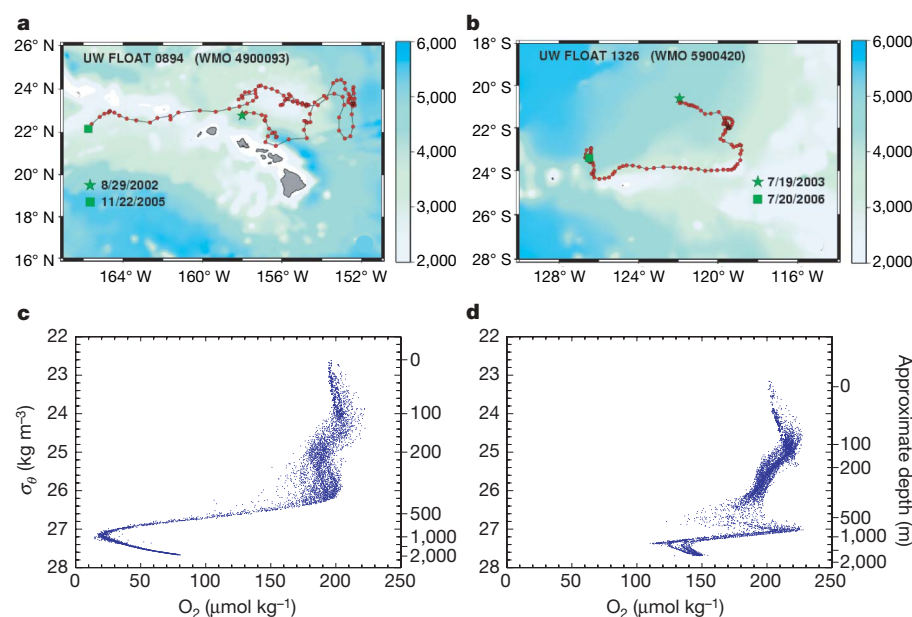
The question of whether the plankton communities in low-nutrient regions of the ocean, comprising 80% of the global ocean surface area, are net producers or consumers of oxygen and fixed carbon is a key uncertainty in the global carbon cycle<sup>1,2</sup>. Direct measurements in bottle experiments indicate net oxygen consumption in the sunlit zone<sup>3–6</sup>, whereas geochemical evidence suggests that the upper ocean is a net source of oxygen<sup>7</sup>. One possible resolution to this conflict is that primary production in the gyres is episodic<sup>1,2,6</sup> and thus difficult to observe: in this model, oligotrophic regions would be net consumers of oxygen during most of the year, but strong, brief events with high primary production rates might produce enough fixed carbon and dissolved oxygen to yield net production as an average over the annual cycle. Here we examine the balance of oxygen production over three years at sites in the North and South Pacific subtropical gyres using the new technique of oxygen sensors deployed on profiling floats. We find that mixing events during early winter homogenize the upper water column and cause low oxygen concentrations. Oxygen then increases below the mixed layer at a nearly constant rate that is similar to independent measures of net community production. This continuous oxygen increase is consistent with an ecosystem that is a net producer of fixed carbon (net autotrophic) throughout the year, with episodic events not required to sustain positive oxygen production.

The uncertainty over whether oligotrophic regions are net producers or consumers of oxygen has “profound implications for our understanding of the oceanic carbon cycle”<sup>7</sup>. However, the direct measurement of oxygen production and respiration in the oligotrophic ocean, which in principle could be used to resolve the

question, is a difficult task. Rates of primary production and respiration are small, with each typically on the order of  $1 \mu\text{mol O}_2 \text{ kg}^{-1} \text{ d}^{-1}$  (ref. 6). Net community production (NCP), which is equal to primary production minus respiration at all trophic levels, is even smaller and more difficult to measure.

We examine here the balance of oxygen production and consumption by using oxygen sensors deployed on two profiling floats<sup>8</sup> as part of the international Argo programme<sup>9</sup> (see Methods). Profiling floats use a buoyancy engine to ascend from a parking depth of 1,000 or 2,000 m every ten days, with oceanographic properties monitored on the ascent and transmitted to shore by satellite. Oxygen measurements are a recent addition to float capabilities<sup>10,11</sup>. Float 0894 collected measurements for three years in the vicinity of the Hawaii Ocean Time series (HOT) station ( $23^\circ \text{N}$ ,  $158^\circ \text{W}$ ; Fig. 1a), whereas float 1326 collected profiles for 3 years near  $22^\circ \text{S}$ ,  $120^\circ \text{W}$  in the South Pacific (Fig. 1b). Oxygen values are plotted against density in Fig. 1c, d for all profiles. The plots illustrate the repeatability of the oxygen measurement, which is  $\pm 1.5 \mu\text{mol kg}^{-1}$  (one standard deviation over 3 years; see Methods). These data are well suited to examining the oxygen balance of subtropical waters because of the relatively rapid ten-day sampling time, high stability and precision of the oxygen measurements, and the fact that these floats remained in nearly homogeneous regions of the ocean for an extended period.

The oxygen concentration and seawater density for the upper 200 m of the water column during the period that each float operated (Fig. 2) indicate that during the autumn of each year the upper 100 m of the water column at the HOT site and  $150 \text{ m}$  at  $22^\circ \text{S}$  undergo strong mixing, homogenizing oxygen and density (Fig. 2). Oxygen concentrations reach low and vertically uniform values during this



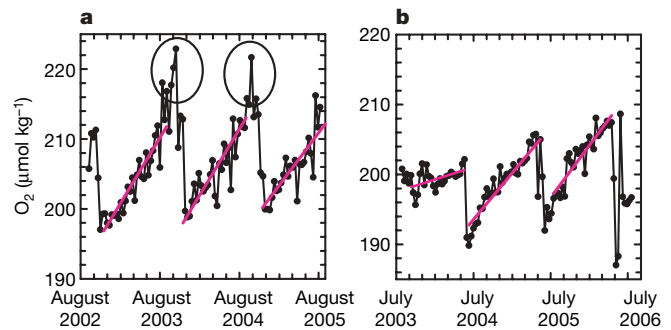
**Figure 1 | Profile locations and oxygen concentration in the subtropical Pacific.** **a, b**, Red dots show the locations of 112 vertical profiles measured by float 0894 in the North Pacific (**a**) and 104 vertical profiles measured by float 1326 in the South Pacific (**b**). Profiles were collected at ten-day intervals. A green star marks the first profile location (29 August 2002 in **a**; 19 July 2003 in **b**) and a green square shows the last profile location (22 November 2005 in **a**; 20 July 2006 in **b**) for each float. The colour bars indicate ocean depth (metres). **c, d**, Raw oxygen data from float 0894 (**c**) and float 1326 (**d**) as functions of potential density  $\sigma_\theta$  and depth, where  $\sigma_\theta$  is adiabatic density  $- 1,000$ . For float 0894, the trajectory is the complete trajectory from launch to the end of the float mission; in the analyses performed here, dissolved oxygen data are used only up to the first 100 profiles of float 0894, after which the oxygen sensor failed.

<sup>1</sup>School of Oceanography, University of Washington, Seattle, Washington 98195, USA. <sup>2</sup>Monterey Bay Aquarium Research Institute, Moss Landing, California 95039, USA.

brief period. For the remainder of each year, oxygen accumulates in the water trapped under the seasonal thermocline. This accumulation of oxygen produces a distinctive shallow oxygen maximum (SOM)<sup>12</sup>.

There is a continuous increase with time in dissolved oxygen in the SOM layer after the autumn period of mixing (Fig. 3). The oxygen anomaly (oxygen concentration minus oxygen solubility) increases at nearly the same rate, indicating that the increase is not driven by solubility changes. The increase in oxygen concentration in the SOM cannot be produced by changes in mixing or solubility and must therefore be due to biological oxygen production, as proposed<sup>12</sup>. The rate of oxygen production was determined from the slopes of straight lines fitted by least squares to the oxygen concentration data from early winter to early autumn (about 300 days) for each year (the pink lines in Fig. 3). Annual NCP rates were estimated from these slopes at depths below the pycnocline by converting oxygen production to carbon uptake with the modified Redfield ratio (150 mol of O<sub>2</sub> produced per 106 mol of CO<sub>2</sub> fixed<sup>13</sup>) and then extrapolating to an annual value by multiplying the daily increase by 365 (Fig. 4).

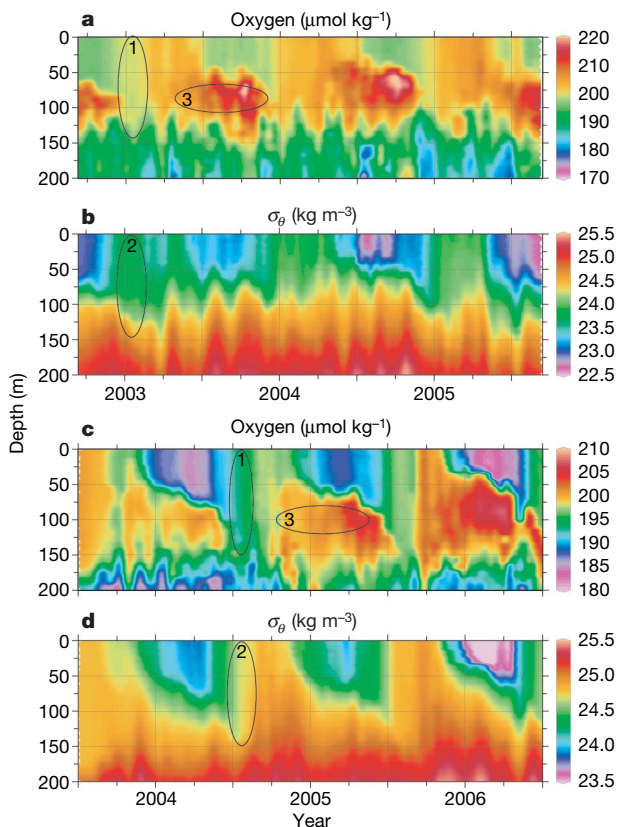
The NCP shows a systematic increase from values not significantly different from zero at depths near the bottom of the euphotic zone (the 1% light level is near 115 m at the HOT site<sup>14</sup> and at about 150 m at 22° S) to maximum values at the base of the pycnocline. The highest slopes are equivalent to a NCP of about 15 mmol C m<sup>-3</sup> yr<sup>-1</sup> near Hawaii and about 7 mmol C m<sup>-3</sup> yr<sup>-1</sup> in the South Pacific gyre. Above the pycnocline, oxygen is lost to the atmosphere by gas exchange, and NCP cannot be reliably estimated from oxygen alone. The yearly cycles of dissolved inorganic carbon in the mixed layer at



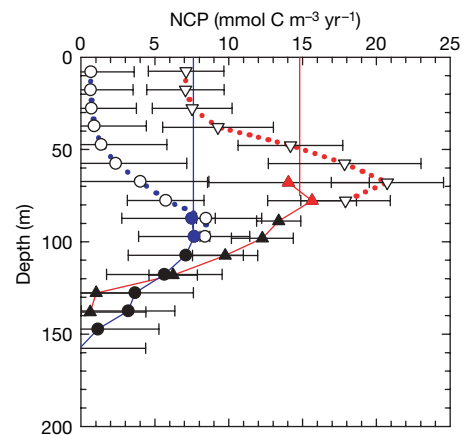
**Figure 3 | Oxygen concentrations in the SOM versus time.** Oxygen concentrations at 78 m for float 0894 (a) and 87 m for float 1326 (b) are shown. Black lines and solid circles are oxygen concentrations measured by the float at each depth. Pink lines are fitted to the oxygen data each year by least squares to estimate the rate of oxygen production. Large black ovals in a identify late summer blooms that increase oxygen concentration in the SOM significantly above the trend line predicted from data earlier in each year.

HOT (Fig. 1 in ref. 15) also have a continuous decrease at rates similar to that of oxygen production. Dissolved inorganic carbon then increases rapidly in the autumn, just as oxygen decreases. Given the similarity of oxygen and dissolved inorganic carbon cycles, the maximum rates at HOT of 15 mmol C m<sup>-3</sup> yr<sup>-1</sup> were extended through the mixed layer to give vertically integrated NCP values of  $1.6 \pm 0.2$  mol C m<sup>-2</sup> yr<sup>-1</sup> (mean  $\pm$  s.d.) (Fig. 4). Keeling *et al.*<sup>15</sup> summarized 11 reported measurements of NCP at the nearby HOT station that had a mean of  $1.9 \pm 0.6$  mol C m<sup>-2</sup> yr<sup>-1</sup> (mean  $\pm$  s.d.). However, Keeling *et al.*<sup>15</sup> reported seasonal variability in NCP that we do not observe. This must have been because they modelled a mean year obtained by averaging 14 years of observations, which obscured the constant rate of NCP.

The reasonable agreement of the NCP derived from float-based oxygen measurements with 11 previous estimates corroborates our hypothesis that the increase in oxygen measured by the floats is the result of biological oxygen production. We conclude that the quasi-lagrangian nature of the floats does not impart a significant bias in the



**Figure 2 | Contours of the evolution of oxygen concentration and density in the upper 200 m.** Oxygen (a, c) and potential density  $\sigma_\theta$  (b, d) are shown during the three years that floats 0894 (a, b) and 1326 (c, d) operated. Contours were prepared with the program Ocean Data View<sup>19</sup>. Periods of convective overturn in 2003 (float 0894) and 2004 (float 1326), during which oxygen and density become vertically homogenous, are identified by ellipses labelled 1 and 2, respectively. The subsequent oxygen increase to form the SOM is identified by ellipses labelled 3.



**Figure 4 | Plot of NCP versus depth.** Triangles show data for float 0894, and circles data for float 1326. Filled symbols were calculated from the slope of oxygen anomaly (oxygen – oxygen solubility) against time in the mixed layer. Open symbols were calculated from the slope of oxygen concentration against time. Vertical solid lines are an extrapolation to the surface of the two highest rates (symbols coloured in red or blue) based on the slope of oxygen against time at each site. Oxygen production was converted to carbon units by using the modified Redfield ratio<sup>13</sup>, as explained in the text. Vertically integrated NCP is the area to the left of the lines connecting filled symbols for each float and the solid line extending that data to the surface. Error bars ( $\pm 1$  s.d.) were computed from the rate of oxygen change for each of the three years for which the floats operated.

measured evolution of oxygen concentration. An estimate of NCP calculated from three years of oxygen data reported by float 1326 near 22° S ( $0.9 \pm 0.4 \text{ mol C m}^{-2} \text{ yr}^{-1}$ ) is about one-half of the HOT value. It is not surprising that the calculated NCP near 22° S is smaller than that near Hawaii (Fig. 4), because float 1326 operated in a region that is considered to have one of the lowest rates of primary production in the world ocean<sup>16</sup>.

The increase in dissolved oxygen beneath the pycnocline follows a relatively smooth trend over time (Fig. 3). Late summer blooms near the HOT site<sup>17</sup> are apparent in the float 0894 data set (Fig. 3a), but they only add to the already positive oxygen increase. The mean rate of increase in oxygen in the SOM near the HOT site is  $0.5 \mu\text{mol kg}^{-1}$  every ten days, excluding the period of the late summer blooms. The observed changes in oxygen in the core of the SOM between each cycle of float 0894 in the months December to June have a frequency distribution that is not significantly different from a normal distribution with a mean of  $0.5 \mu\text{mol kg}^{-1}$  and a standard deviation of  $1.5 \mu\text{mol kg}^{-1}$  (Kolmogorov–Smirnov test;  $P = 0.20, 0.11$  and  $0.89$  for 2003, 2004 and 2005, respectively). Such a distribution is consistent with a constant rate of increase in oxygen along with the analytical precision determined from deep oxygen measurements. If the data set is extended to September, then the Kolmogorov–Smirnov test fails ( $P = 0.06$  for 2003 and 2004, with no August or September data in 2005) because of the late summer blooms. This confirms that we do detect episodic events when they are present. There is no evidence that episodic events at a frequency lower than the float cycle contribute to the oxygen increase before the late summer blooms. Aperiodic increases in oxygen concentration have been reported over 2–3-month intervals with oxygen sensors at 50 m depth, near the top of the SOM, on a mooring deployed near the HOT site<sup>18</sup>. Similar variability is produced in the data from float 0894 by small vertical excursions in the sharp oxygen gradient at the top of the SOM (Fig. 2a), as documented by simultaneous density variations (Fig. 2b); these are not episodic production events. If there is an episodic component to NCP, it must occur at intervals that are equal to or shorter than the float cycle period. Such a process would be more nearly periodic than episodic. We conclude that the float oxygen data provide unambiguous evidence that the euphotic zones in the North and South Pacific subtropical gyres are net producers of oxygen. Infrequent, episodic events are not required to sustain positive NCP.

## METHODS SUMMARY

The data were collected with Webb Research Apex profiling floats constructed at the University of Washington. These floats were parked at 1,000 m depth and ascended to the surface at ten-day intervals. Seabird SBE43 sensors measured oxygen concentrations at 50 depths during the ascent. The oxygen data analysed here consist of the raw, transmitted values and have not been adjusted in any way.

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Received 26 February; accepted 5 November 2007.**

1. del Giorgio, P. A. & Duarte, C. M. Respiration in the open ocean. *Nature* **420**, 379–384 (2002).

2. Karl, D. M., Laws, E. A., Morris, P., Williams, P. J., le B. & Emerson, S. Metabolic balance in the open sea. *Nature* **426**, 32 (2003).
3. del Giorgio, P. A., Cole, J. J. & Cimbleris, A. Respiration rates in bacteria exceed phytoplankton production in unproductive aquatic systems. *Nature* **385**, 148–151 (1997).
4. Duarte, C. M. & Agusti, S. The CO<sub>2</sub> balance of unproductive aquatic ecosystems. *Science* **281**, 234–236 (1998).
5. Duarte, C. M., Agusti, S., del Giorgio, P. A. & Cole, J. J. Regional carbon imbalances in the oceans. *Science* **284**, 173–174 (1999).
6. Williams, P. J., le B., Morris, P. J. & Karl, D. M. Net community production and metabolic balance at the oligotrophic ocean site, station ALOHA. *Deep-Sea Res. I* **51**, 1563–1578 (2004).
7. Williams, P. J., le B. & Bowers, D. G. Regional carbon imbalances in the oceans. *Science* **284**, 173–174 (1999).
8. Roemmich, D., Riser, S., Davis, R. & Desaubies, Y. Autonomous profiling floats: workhorse for broad-scale observations. *Mar. Technol. Soc. J.* **38**, 21–29 (2004).
9. Roemmich, D. *et al.* in *Observing the Oceans in the 21st Century* (eds Koblenz, K. & Smith, N.) 248–258 (Australian Bureau of Meteorology, Melbourne, Australia, 2001).
10. Kortzinger, A., Schimanski, J., Send, U. & Wallace, D. The ocean takes a deep breath. *Science* **306**, 1337 (2004).
11. Johnson, K. S., Needoba, J. A., Riser, S. C. & Showers, W. J. Chemical sensor networks for the aquatic environment. *Chem. Rev.* **107**, 623–640 (2007).
12. Schulenberger, E. & Reid, J. L. The Pacific shallow oxygen maximum, deep chlorophyll maximum, and primary productivity reconsidered. *Deep-Sea Res. A* **28**, 901–919 (1981).
13. Anderson, L. A. On the hydrogen and oxygen content of marine phytoplankton. *Deep-Sea Res. I* **42**, 1675–1680 (1995).
14. Letelier, R. M., Karl, D. M., Abbott, M. R. & Bidigare, R. R. Role of late winter mesoscale events in the biogeochemical variability of the upper water column of the North Pacific Subtropical Gyre. *J. Geophys. Res.* **105**, 28723–28739 (2000).
15. Keeling, C. D., Brix, H. & Gruber, N. Seasonal and long-term dynamics of the upper ocean carbon cycle at Station ALOHA near Hawaii. *Glob. Biogeochem. Cycles* **18**, doi:10.1029/2004GB002227 (2004).
16. Behrenfeld, M. J., Boss, E., Siegel, D. A. & Shea, D. M. Carbon-based ocean productivity and phytoplankton physiology from space. *Glob. Biogeochem. Cycles* **19**, doi:10.1029/2004GB002299 (2005).
17. Scharek, R., Tupas, L. M. & Karl, D. M. Diatom fluxes to the deep sea in the oligotrophic North Pacific gyre at Station Aloha. *Mar. Ecol. Prog. Ser.* **182**, 55–67 (1999).
18. Emerson, S., Stump, C., Johnson, B. & Karl, D. M. In situ determination of oxygen and nitrogen dynamics in the upper ocean. *Deep-Sea Res. I* **49**, 941–952 (2002).
19. Schlitzer, R. Ocean Data View (<http://odv.awi.de>) (2006).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank N. Larson for producing the oxygen sensors; D. Swift for his essential contributions to this effort; and the Hawaii Ocean Time-series participants for making dissolved oxygen data available. Research at the University of Washington was supported through the US Argo Program by the National Oceanographic and Atmospheric Administration and by the US Office of Naval Research through the National Ocean Partnership Program. Research at Monterey Bay Aquarium Research Institute was supported by a grant from the David and Lucile Packard Foundation and by the National Science Foundation.

**Author Contributions** S.C.R. originated the idea of putting oxygen sensors on profiling floats, and directed the construction and deployment of the floats as part of the international Argo project. K.S.J. performed the data analysis. Both authors contributed to the writing of the manuscript.

**Author Information** All float data are available from the global Argo data center at <ftp://usgodaefnmoc.navy.mil/pub/outgoing/argo/>. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence and requests for materials should be addressed to S.C.R. ([riser@ocean.washington.edu](mailto:riser@ocean.washington.edu)) or K.S.J. ([johnson@mbari.org](mailto:johnson@mbari.org)).

## METHODS

Nearly 3,000 profiling floats are now deployed in the ocean as part of Argo<sup>9</sup>, and about 70 of them have been equipped with sensors for dissolved oxygen<sup>10,11</sup>. Here we use oxygen concentrations that were measured over the course of three years by University of Washington floats 0894 (WMO Number 4900093) and 1326 (WMO Number 5900420). Oxygen concentration, temperature, salinity and pressure were measured on the ascent at 50 depths between 1,000 m and 7 m, and the data were transmitted to shore while the floats were at the surface. On every fourth profile, the floats descended to 2,000 m depth before profiling to the surface, and measurements were made at 70 depths to the surface. The floats then descended to their parking depth and drifted before returning to the surface to repeat the cycle at ten-day intervals.

Oxygen concentrations were measured on both floats with Seabird SBE43 sensors for dissolved oxygen. The SBE43 oxygen sensor is in the flow stream that is fed by the conductivity sensor pump, where it is protected from biofouling by the Seabird conductivity-cell antifouling system. This results in exceptional long-term stability that can be quantified from the variability of oxygen concentrations measured in the deep waters sampled over three years on the profiles from 2,000 m (Supplementary Fig. 1a, b). During the ten-day interval between profiles, oxygen is depleted around the sensor because it is always polarized and consuming oxygen. The oxygen sensor on float 1326 did not have sufficient time, after pump turn-on, to respond to the ambient oxygen concentration, and its initial measurements at 2,000 m were biased low. Stability of the sensor on float 1326 was quantified at 1,900 m depth. We assessed sensor performance on float 894 at 1,800 m depth, which coincides with a standard depth for oxygen measurements at the HOT time series station.

The concentration of oxygen measured by float 894 at 1,800 m over three years averages  $70.8 \pm 1.3 \mu\text{mol kg}^{-1}$  (mean  $\pm$  s.d.,  $n = 25$ ; Supplementary Fig. 1a). Equivalent precision is obtained for data from 2,000 m. During the same period, oxygen concentrations measured by Winkler titration in discrete samples from 1,800 m in the HOT time series averaged  $79.0 \pm 2.1 \mu\text{mol kg}^{-1}$  ( $n = 31$ ; Supplementary Fig. 1a). The oxygen measured at 1,900 m by float 1326 averaged  $145.9 \pm 1.5 \mu\text{mol kg}^{-1}$  ( $n = 26$ ) over three years (Supplementary Fig. 1b). There was a shift in water mass structure at shallower depths, as float 1326 drifted south, resulting in a bifurcation in the plot of deep oxygen against density (Fig. 1d). This same bifurcation is seen in the temperature–salinity plot (not shown). It is also seen in data from the same region on the World Ocean Circulation Experiment P17 survey, which was a meridional section along

135° W. The results for both sensors demonstrate that the precision (1 s.d.) and stability of the dissolved oxygen measurements over three years are better than  $1.5 \mu\text{mol kg}^{-1}$ .

As further evidence of the stability of the sensors, we consider data for oxygen concentration against time at 200 m depth, which lies below the euphotic zone. During the three-year period over which float 0894 operated (Supplementary Fig. 1c) the change in oxygen at 200 m was smaller than  $-0.1 \mu\text{mol kg}^{-1} \text{yr}^{-1}$ . Float 1326 did detect a slight upward trend over time ( $1.9 \mu\text{mol kg}^{-1} \text{yr}^{-1}$ ) at 200 m (Supplementary Fig. 1d), which was paralleled by an equivalent trend in oxygen solubility<sup>20</sup> as the float drifted south; again, the sensor appeared stable. However, the variability in oxygen concentrations at 200 m depth was much larger than that found at depths below 1,000 m (Supplementary Fig. 1a, b) or near the surface (Supplementary Fig. 2). Oxygen concentrations determined by Winkler titration at 200 m in the HOT time series over the same period ( $197 \pm 6 \mu\text{mol kg}^{-1}$ ) have similar variability to the float data ( $188 \pm 6 \mu\text{mol kg}^{-1}$ ; Supplementary Fig. 1c). The high variability at 200 m must reflect real changes in oxygen concentration driven by physical and biological processes, and it does not reflect sensor precision.

Variability in oxygen concentration within the mixed layer primarily reflects changes in oxygen solubility (Supplementary Fig. 2a, b). As the water column warms during spring and summer, oxygen solubility<sup>20</sup> decreases and the oxygen in the mixed layer outgasses to the atmosphere. Although the float data suggest that concentration of oxygen in the mixed layer averages  $10 \mu\text{mol kg}^{-1}$  (float 0894) and  $16 \mu\text{mol kg}^{-1}$  (float 1326) below the concurrent solubility values (Fig. 2a, b), these differences probably reflect inaccuracies in the absolute calibration of the oxygen sensors. Measurements of near-surface oxygen generally indicate a supersaturation of 0 to  $+4 \mu\text{mol kg}^{-1}$ , particularly near the HOT site<sup>18</sup>. Similar offsets in the calibration of the sensor on float 0894 are seen at 200 and 1,800 m when compared with the Winkler titration data at the HOT site (Supplementary Figs 1c and 2a).

Despite the calibration offsets, the inherent noise and drift of the oxygen sensors are less than one-tenth of the annual changes in oxygen detected in the SOM. To exploit this high precision, here we concern ourselves with relative changes in the concentration of oxygen over time rather than the absolute accuracy of the oxygen sensor.

20. Weiss, R. F. The solubility of nitrogen, oxygen and argon in water and seawater. *Deep-Sea Res. A* 17, 721–735 (1970).

## LETTERS

# Dry mantle transition zone inferred from the conductivity of wadsleyite and ringwoodite

Takashi Yoshino<sup>1</sup>, Geeth Manthilake<sup>1</sup>, Takuya Matsuzaki<sup>1</sup> & Tomoo Katsura<sup>1</sup>

The Earth's mantle transition zone could potentially store a large amount of water, as the minerals wadsleyite and ringwoodite incorporate a significant amount of water in their crystal structure<sup>1,2</sup>. The water content in the transition zone can be estimated from the electrical conductivities of hydrous wadsleyite and ringwoodite, although such estimates depend on accurate knowledge of the two conduction mechanisms in these minerals (small polaron and proton conduction), which early studies have failed to distinguish between<sup>3,4</sup>. Here we report the electrical conductivity of these two minerals obtained by high-pressure multi-anvil experiments. We found that the small polaron conduction of these minerals are substantially lower than previously estimated. The contributions of proton conduction are small at temperatures corresponding to the mantle transition zone and the conductivity of wadsleyite is considerably lower than that of ringwoodite for both mechanisms. The dry model mantle shows considerable conductivity jumps associated with the olivine–wadsleyite, wadsleyite–ringwoodite and post-spinel transitions. Such a dry model explains well the currently available conductivity–depth profiles<sup>5</sup> obtained from geoelectromagnetic studies. We therefore conclude that there is no need to introduce a significant amount of water in the mantle transition to satisfy electrical conductivity constraints.

Electrical conductivity is useful in studying the composition, mineralogy and temperature of the Earth's deep interior. The electrical conductivity of the mantle constituent minerals is mostly influenced by proton ( $H^+$ ) and small polaron conduction (electron holes hopping between  $Fe^{2+}$  and  $Fe^{3+}$ )<sup>6</sup> mechanisms. In other words, conductivity is sensitive to small amounts of hydrogen<sup>7</sup> and iron. Therefore, to estimate the water content of the mantle transition zone, the contributions of small polaron and proton conduction must be separately determined to reach a full understanding of the electrical conductivity of wadsleyite and ringwoodite. Xu *et al.*<sup>3</sup> reported that the electrical conductivities of wadsleyite and ringwoodite are similar and two orders of magnitude higher than that of olivine. However, their conductive values are too high to explain the recent conductivity–depth profiles in the transition zone obtained by semi-global electromagnetic induction studies<sup>5,8–11</sup>. Although Xu *et al.*<sup>3</sup> considered that small polaron conduction was the dominant conduction mechanism in their study, Huang *et al.*<sup>4</sup> later attributed the results to proton conduction because they found a significant amount of water in the samples that Xu *et al.*<sup>3</sup> used. This means that we have no data about the small polaron conduction of these minerals. Although Huang *et al.*<sup>4</sup> claimed that they determined the proton conduction of these minerals, their results can be considered invalid because of serious methodological problems (see Supplementary Information). Thus, at present, we have no understanding of either small polaron or proton conduction in these minerals. Here we determine the conductivities of wadsleyite and

ringwoodite by distinguishing between small polaron and proton conduction mechanisms.

The electrical conductivities of wadsleyite and ringwoodite were measured in a Kawai-type multi-anvil apparatus over several heating–cooling cycles (see Supplementary Information). To clarify the effect of hydrogen on the electrical conductivity, we have conducted conductivity measurements both for initially hydrogen-doped samples and for undoped samples. Based on the technique of ref. 12, conductivity measurements were made using low-frequency (0.1–0.01 Hz) alternating current signals in a temperature range from 300 to 2,000 K and pressure conditions at 16 GPa for wadsleyite and 20 GPa for ringwoodite. Complex impedance spectroscopic analyses were also carried out over a wide frequency range (1 MHz to 0.01 Hz) to confirm the validity of the low-frequency data (see Supplementary Information). For the hydrogen-doped samples, conductivity was measured under lower temperature conditions (<1,000 K) to minimize water loss. The samples were characterized by X-ray diffraction, electron microprobe analysis and electron microscopic observation. The water content of the samples was determined by non-polarized Fourier-transform infrared spectroscopy both before and after each conductivity measurement. The Paterson calibration was used to calculate the water content from the infrared absorption<sup>13</sup> (detailed in the Supplementary Information).

Figure 1 shows an Arrhenius plot showing the conductivity of the hydrogen-doped and undoped wadsleyite and ringwoodite containing various amounts of water. The water content (in wt%) detected for each sample is also shown. The hydrogen-undoped samples, which experienced temperatures higher than 1,700 K, contained measurable amounts of water, which is considered to come from the surrounding pressure medium at high temperatures, as was the case for olivine<sup>6</sup>. In contrast, we found no change of water content during the conductivity measurement for the hydrogen-doped samples, which experienced temperatures only up to 1,000 K (see Supplementary Fig. 4). The absolute conductivity values increase with increasing water content especially at low temperatures, suggesting that proton conduction dominates at lower temperatures. The temperature dependence at low temperatures decreases with increasing water content, which is particularly the case for ringwoodite. In contrast, the conductivity is relatively independent of the water content at high temperatures, where small polaron conduction is considered to be dominant. The results using the pre-synthesized and hydrogen-undoped samples (relatively dry: <100 weight p.p.m.  $H_2O$ ) are consistent with those using the olivine single crystal as a sample. In any temperature range, the absolute electrical conductivity of wadsleyite is lower than that of ringwoodite at the same temperature and water content. For both minerals, conductivity in the high-temperature region is significantly lower than that obtained previously<sup>3</sup>.

<sup>1</sup>Institute for Study of the Earth's Interior, Okayama University, Misasa, Tottori 682-0193, Japan.

The electrical conductivity ( $\sigma$ ) of a hydrous iron-bearing silicate mineral can be expressed by the following equation:

$$\sigma = \sigma_{0H} \exp\left(-\frac{H_H}{kT}\right) + \sigma_{0P} \exp\left(-\frac{H_P}{kT}\right) \quad (1)$$

where  $\sigma_0$  is the pre-exponential factor,  $H$  is the activation enthalpy,  $k$  is the Boltzmann constant and  $T$  is temperature. Subscripts H and P denote small polaron (hopping) and proton conduction, respectively. At low temperatures, the small polaron conduction is masked by the proton conduction because of the smaller temperature dependence of proton conduction. For ringwoodite, the apparent activation enthalpy at lower temperatures decreases from 1.1 to

0.5 eV with increasing water content. The decrease in activation energy ( $H_A$ ) could be caused by a change of dominant hydrogen configuration in the crystal structure with increasing water content (see Supplementary Information) and be approximated well by an equation similar to that for an n-type semiconductor<sup>14</sup>:

$$H_A(N_A^-) = H_A(0) - \alpha(N_A^-)^{1/3} \quad (2)$$

where  $N_A^-$  is the hydrogen concentration in the crystal structure (number of atoms per unit volume),  $H_A(N_A^-)$  is the value of  $H_A$  at a certain value of  $N_A^-$ ,  $H_A(0)$  is the activation energy observed at very low hydrogen concentrations, and  $\alpha$  is a constant accounting for geometrical factors (Supplementary Fig. 5). The pre-exponential factor ( $\sigma_{0H}$ ) in equation (1) is generally defined as a function of water content<sup>7</sup>. It is known from the Nernst–Einstein equation that electrical conductivity depends on the number ( $N$ ) of electric charge carriers per unit volume:

$$\sigma = Nze\mu \quad (3)$$

where  $z$  is the charge number (for a proton  $z = 1$ ),  $e$  is the charge of an electron and  $\mu$  is mobility (hydrogen diffusion).

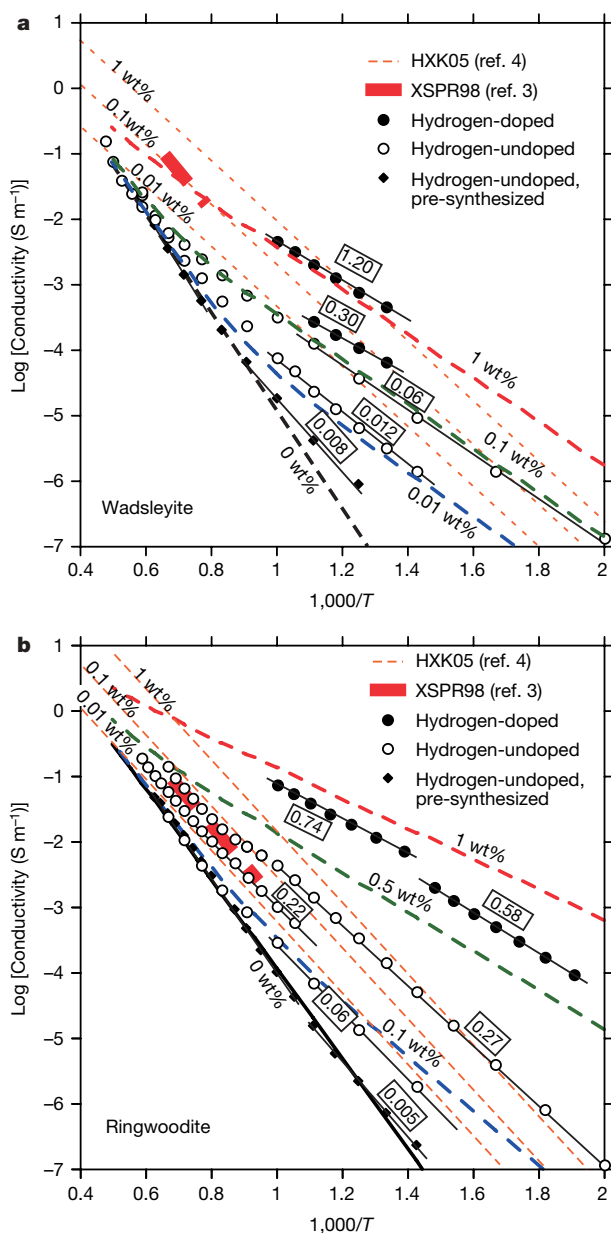
Taking into account the concentration dependence of the pre-exponential factor, the resultant electrical conductivities can be expressed as follows:

$$\sigma = \sigma_{0H} \exp\left(-\frac{H_H}{kT}\right) + \sigma_{0P} C_w \exp\left(-\frac{H_P^0 - \alpha C_w^{1/3}}{kT}\right) \quad (4)$$

where  $C_w$  is water content in weight percent. The fitting parameters for wadsleyite and ringwoodite are summarized in Table 1. The activation energies of wadsleyite and ringwoodite for small polaron conduction are 1.49 and 1.36 eV, respectively. These values are comparable with those for small polaron conduction in olivine (~1.3–1.6 eV)<sup>3,6,15</sup>. The dependence on water content of activation energy for proton conduction in ringwoodite is large, whereas that in wadsleyite is negligibly small (see Supplementary Information). For wadsleyite, the activation energy (~0.7 eV) for proton conduction is distinctly lower than that (1.27 eV) obtained from hydrogen diffusion experiments<sup>16</sup>, as is the case for olivine<sup>6</sup>.

To assess the presence of water in the mantle transition zone, we compared a laboratory-based conductivity–depth profile constructed using the present experimental data with that obtained from electromagnetic studies. We made the following assumptions to construct the profile. The effects of activation volume, grain size, the presence of additional phases (such as majorite) and microstructure on the bulk conductivity were not considered. The oxygen fugacity in the mantle transition zone was taken to be that for iron-wüstite<sup>17,18</sup>. Because the iron-wüstite buffer is probably close to the Mo/MoO<sub>2</sub> buffer<sup>3</sup>, our experimental data were used directly. The geotherm (1,780–1,925 K in the transition zone) was assumed to be adiabatic, and was taken from Katsura *et al.*<sup>19</sup>. We used previous conductivity data for olivine<sup>6,15</sup> and perovskite<sup>20</sup> to estimate conductivity above and below the transition zone. To calculate the conductivity–depth profile down to the 410-km discontinuity for small polaron conduction, we used data for single-crystal olivine on a Ni–NiO buffer<sup>6</sup> and for polycrystalline olivine on a Mo–MoO<sub>2</sub> buffer<sup>15</sup>. For the single-crystal data, we used the average value of the three axes at any given temperature.

Figure 2 shows a profile of conductivity versus depth as a function of water content in a range from 200 to 800 km depth based on our experimental data. The conductivity values in the dry transition zone presented here are significantly lower than those estimated from



**Figure 1 | Electrical conductivity of wadsleyite and ringwoodite as a function of reciprocal temperature.** **a**, Wadsleyite; **b**, ringwoodite. The symbols indicate raw data for each sample with different water contents. Previous results from Xu *et al.*<sup>3</sup> and Huang *et al.*<sup>4</sup> are shown as a function of water content. Coloured thick dashed lines indicate the electrical conductivity calculated by data fitting based on equation (4) as a function of water content. Numbered boxes denote the estimated water content (in weight percent) by Fourier-transform infrared analysis. Errors for the estimated water content become larger with decreasing water content and range from  $\pm 20$  (~1 wt%) to  $\pm 50\%$  (<0.01 wt%).

**Table 1 | Parameter values**

Mineral	$\sigma_{0H}$ (S m <sup>-1</sup> )	$H_H$ (eV)	$\sigma_{0P}$ (S m <sup>-1</sup> )	$H_P^0$ (eV)	$\alpha$
Wadsleyite	399(311)	1.49(10)	7.74(4.08)	0.68(3)	0.02(2)
Ringwoodite	838(442)	1.36(5)	27.7(9.6)	1.12(3)	0.67(3)

Numbers in parentheses are the errors by nonlinear least squares fitting (1 $\sigma$  standard deviation).

ref. 3. The present model shows three distinct conductivity jumps at depths of 410 km (0.3 and 0.7 log units for Ni–NiO and Fe–FeO buffers, respectively), 520 km (0.8 log units for the wadsleyite–ringwoodite transition) and 660 km (0.5 log units for the post-spinel transition) without water. In contrast, the previous model<sup>3</sup> showed a fairly large conductivity jump at 410 km depth and a negligibly small jump at 520 km. In our model, the magnitude of the conductivity jump at the 410-km discontinuity decreases and the one at 520 km increases with increasing water content. If wadsleyite and ringwoodite hold 1 wt% water, conductivity increases by ~1 order of magnitude compared with the dry mantle model. It is difficult to determine water content less than 0.1 wt% for the normal geotherm because the contribution of proton conduction, with its low activation enthalpies, will be hidden by small polaron conduction with high activation enthalpy at such high temperatures of the mantle transition zone.

Utada *et al.*<sup>10</sup> constructed a one-dimensional electrical conductivity profile beneath the North Pacific Ocean given by semi-global electromagnetic induction studies covering one-quarter of the Earth. They used the data from eight submarine cables combined with data from 17 geomagnetic observatories, which is considered to be the best available data set at present. Kuvshinov *et al.*<sup>5</sup> later reanalysed the same data set by means of much more sophisticated correction using the detailed three-dimensional ocean model, producing the most reliable conductivity profile so far available. This model<sup>5</sup> suggests that the oceanic mantle in the transition zone is much more resistive than previously proposed<sup>10</sup>.

Our laboratory-based conductivity model of the dry mantle transition zone explains the conductivity profile of ref. 5 very well. These two models are also in excellent agreement above the transition zone if the oxygen fugacity is that obtained with the Mo–MoO<sub>2</sub> buffer. In addition, the conductivity jump at the 660-km discontinuity agrees with the conductivity difference between perovskite and ringwoodite. In contrast, the previous mineralogical model<sup>3</sup> cannot explain this conductivity profile<sup>5</sup>. Although Xu *et al.*<sup>3</sup> justified their results

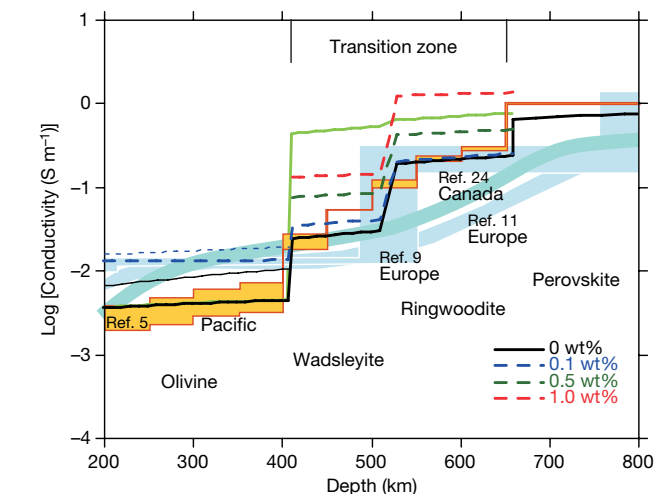
based on past studies proposing a significant conductivity jump at 400 km depth<sup>21,22</sup>, other recent one-dimensional models<sup>9,11,23,24</sup> also show no evidence of a conductivity jump of two orders of magnitude related to the 410-km discontinuity.

The conductivity profile shown in ref. 5 was constructed without consideration of the wadsleyite–ringwoodite transition, so the conductivity jump at the 520-km discontinuity is missing from this profile. In addition, beneath continents, electrical conductivities below the 520-km discontinuity (ringwoodite stability field) are slightly lower than those beneath the ocean (around  $10^{-2}$  S m<sup>-1</sup>). However, the conductivity–depth profile beneath Europe produced by the layered model of ref. 9 has a large discontinuity in the middle of the transition zone rather than its top and bottom, which would suggest the presence of a conductivity jump at the wadsleyite–ringwoodite transition. We suggest that an electromagnetic study should construct conductivity profiles by assuming conductivity jumps associated with all the major phase transitions.

As shown in Fig. 2, conductivity above the transition zone is considerably higher beneath continents than beneath the ocean<sup>5,8,9,11,24</sup>. As a result, no clear conductivity jump is seen at the 410-km discontinuity beneath continents. The large conductivity values above the transition zone and small conductivity jump at 410 km depth are well explained if a more oxidized (Ni–NiO buffer) condition is assumed. Beneath the French Alps, the electrical conductivity values in the transition zone are low, around  $10^{-2}$  S m<sup>-1</sup> (ref. 11). The normal geotherm model cannot account for such conductivity values. This had been explained by proposing that the dry subducting slab is cooler than the surrounding mantle. Although Tarits *et al.*<sup>11</sup> estimated a temperature of 350–450 K less than the normal geotherm based on previous experimental results<sup>3</sup>, our study can explain these values by proposing a temperature reduction of only 150 K relative to the normal geotherm. Beneath the Canadian Shield, the conductivity values in the transition zone are also lower than beneath the ocean, especially in the ringwoodite stability field. If the region has the normal geotherm, a presence of iron-poor ringwoodite might be expected.

Our conductivity model explains the conductivity–depth profiles of the oceanic and continent mantle in the transition zone very well, without the contribution of water. Therefore, there is no need to incorporate water into wadsleyite and ringwoodite. On the other hand, the possibility of less than 0.1% of water cannot be excluded, because of the relatively small contribution of proton conduction at high temperatures. Electrical conductivity can be used to estimate the water content in the mantle if it is above 0.1 wt%.

Received 29 May; accepted 18 October 2007.



**Figure 2 | Electrical conductivity profiles beneath the Pacific, and the estimated water content in the mantle transition zone.** The orange and bluish areas represent geophysically observed conductivity profiles in the Pacific from ref. 5 and the continental mantle from refs 9, 11 and 24, respectively. The thick solid line represents the electrical conductivity of olivine, wadsleyite and ringwoodite without water. Dashed lines indicate the electrical conductivity of hydrous olivine, wadsleyite and ringwoodite as a function of water content (red: 1.0 wt%; green: 0.5 wt%; blue: 0.1 wt%). Light green solid line denotes the previous experimental result of Xu *et al.*<sup>3</sup>. The electrical conductivity of hydrous olivine was estimated from the average of three crystallographic axes<sup>6</sup>. In the olivine stability field, thick and thin lines indicate the electrical conductivity estimated from different conditions of oxygen fugacity, with Mo–MoO<sub>2</sub> and Ni–NiO buffers, respectively.

1. Inoue, T. Effect of water on melting phase relations and melt composition in the Mg<sub>2</sub>SiO<sub>4</sub>–MgSiO<sub>3</sub>–H<sub>2</sub>O system up to 15 GPa. *Phys. Earth Planet. Inter.* **85**, 237–263 (1994).
2. Kohlstedt, D. L., Keppler, H. & Rubie, D. C. Solubility of water in the  $\alpha$ ,  $\beta$  and  $\gamma$  phases of (Mg,Fe)<sub>2</sub>SiO<sub>4</sub>. *Contrib. Mineral. Petrol.* **123**, 345–357 (1996).
3. Xu, Y., Shankland, T., Poe, B. & Rubie, D. C. Electrical conductivity of olivine, wadsleyite and ringwoodite under upper-mantle condition. *Science* **280**, 1415–1418 (1998).
4. Huang, X., Xu, Y. & Karato, S. Water content in the transition zone from electrical conductivity of wadsleyite and ringwoodite. *Nature* **434**, 746–749 (2005).
5. Kuvshinov, A., Utada, H., Avdeev, A. & Koyama, T. 3-D modelling and analysis of Dst C-responses in the North Pacific Ocean region, revisited. *Geophys. J. Int.* **160**, 505–526 (2005).
6. Yoshino, T., Matsuzaki, T., Yamashita, S. & Katsura, T. Hydrous olivine unable to account for conductivity anomaly at the top of the asthenosphere. *Nature* **443**, 973–976 (2006).
7. Karato, S. The role of hydrogen in the electrical conductivity of the upper mantle. *Nature* **347**, 272–273 (1990).
8. Schultz, A., Kurtz, R. D., Chave, A. D. & Jones, A. D. Conductivity discontinuities in the upper mantle beneath a stable craton. *Geophys. Res. Lett.* **20**, 2941–2944 (1993).
9. Olsen, N. The electrical conductivity of the mantle beneath Europe derived from C-responses from 3 to 720 hr. *Geophys. J. Int.* **133**, 298–308 (1998).
10. Utada, H., Koyama, T., Shimizu, H. & Chave, A. D. A semi-global reference model for electrical conductivity in the mid-mantle beneath the north Pacific region. *Geophys. Res. Lett.* **30**, 1194, doi:10.1029/2002GL016902 (2003).

11. Tarits, P., Hautot, S. & Perrier, F. Water in the mantle: Results from electrical conductivity beneath the French Alps. *Geophys. Res. Lett.* **31**, L06612, doi:10.1029/2003GL019277 (2004).
12. Fu-jita, K., Katsura, T. & Tainosho, Y. Electrical conductivity measurement of granulite under mid- to lower crustal pressure-temperature conditions. *Geophys. J. Int.* **157**, 79–86 (2004).
13. Patterson, M. S. The determination of hydroxyl by infrared absorption in quartz, silicate glasses and similar minerals. *Bull. Mineral.* **105**, 20–29 (1982).
14. Debye, P. P. & Conwell, E. M. Electrical properties of N-type germanium. *Phys. Rev.* **93**, 693–706 (1954).
15. Xu, Y., Shankland, T. J. & Duba, A. G. Pressure effect on electrical conductivity of mantle olivine. *Phys. Earth Planet. Inter.* **118**, 149–161 (2000).
16. Hae, R., Ohtani, E., Kubo, T., Koyama, T. & Utada, H. Hydrogen diffusivity in wadsleyite and water distribution in the mantle transition zone. *Earth Planet. Sci. Lett.* **243**, 141–148 (2006).
17. McCammon, C. The paradox of mantle redox. *Science* **308**, 807–808 (2005).
18. Hirschmann, M. A wet mantle conductor? *Nature* **439**, E3–E4, doi:10.1038/nature04529 (2006).
19. Katura, T. *et al.* Olivine-wadsleyite transition in the system (Mg,Fe)<sub>2</sub>SiO<sub>4</sub>. *J. Geophys. Res.* **109**, B02209, doi:10.1029/2003JB002438 (2004).
20. Katsura, T., Sato, K. & Ito, E. Electrical conductivity of silicate perovskite at lower-mantle condition. *Nature* **395**, 493–495 (1998).
21. Banks, R. J. Geomagnetic variations and the electrical conductivity of the mantle. *Geophys. J. R. Astron. Soc.* **17**, 457–487 (1969).
22. Bahr, K., Olsen, N. & Shankland, T. J. On the combination of the magnetotelluric and the geomagnetic depth sounding method for resolving of an electrical conductivity increase at 400 km depth. *Geophys. Res. Lett.* **20**, 2937–2940 (1993).
23. Lizzarralde, D., Chave, A. D., Hirth, G. & Schultz, A. Northeastern Pacific mantle conductivity profile from long-period magnetotelluric sounding using Hawaii to California submarine cable data. *J. Geophys. Res.* **100**, 17837–17854 (1995).
24. Neal, S. L., Mackie, R. L., Larsen, J. C. & Schultz, A. Variations in the electrical conductivity of the upper mantle beneath North America and the Pacific Ocean. *J. Geophys. Res.* **105**, 8229–8242 (2000).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank E. Ito, D. Yamazaki for critical discussion, S. Yamashita and N. Bolfan-Casanova for interpretation of Fourier-transform infrared spectra, H. Utada for beneficial discussion of conductivity structure and C. Oka for technical assistance. This research was supported by a Grant-in-Aid for Scientific Research to T.K. and T.Y. from the Japan Society for the Promotion of Science and the COE-21 program to the Institute for Study of the Earth's Interior, Okayama University.

**Author Contributions** T.K. and T.Y. organized the project and completed the manuscript. The conductivity measurements of wadsleyite and ringwoodite were made by G.M. and T.Y., respectively. The Fourier-transform infrared analysis was made by T.M.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence and requests for materials should be addressed to T.Y. (tyoshino@misasa.okayama-u.ac.jp).

## LETTERS

# Reversal of pathological pain through specific spinal GABA<sub>A</sub> receptor subtypes

Julia Knabl<sup>1</sup>, Robert Witschi<sup>2</sup>, Katharina Hösl<sup>1</sup>, Heiko Reinold<sup>1</sup>, Ulrike B. Zeilhofer<sup>1</sup>, Seifollah Ahmadi<sup>1†</sup>, Johannes Brockhaus<sup>2†</sup>, Marina Sergejeva<sup>1</sup>, Andreas Hess<sup>1</sup>, Kay Brune<sup>1</sup>, Jean-Marc Fritschy<sup>2</sup>, Uwe Rudolph<sup>2,4</sup>, Hanns Möhler<sup>2,3,5</sup> & Hanns Ulrich Zeilhofer<sup>1,2,3</sup>

Inflammatory diseases and neuropathic insults are frequently accompanied by severe and debilitating pain, which can become chronic and often unresponsive to conventional analgesic treatment<sup>1,2</sup>. A loss of synaptic inhibition in the spinal dorsal horn is considered to contribute significantly to this pain pathology<sup>3–7</sup>. Facilitation of spinal  $\gamma$ -aminobutyric acid (GABA)<sub>A</sub> receptor neurotransmission through modulation of GABA<sub>A</sub> receptors should be able to compensate for this loss<sup>8,9</sup>. With the use of GABA<sub>A</sub>-receptor point-mutated knock-in mice in which specific GABA<sub>A</sub> receptor subtypes have been selectively rendered insensitive to benzodiazepine-site ligands<sup>10–12</sup>, we show here that pronounced analgesia can be achieved by specifically targeting spinal GABA<sub>A</sub> receptors containing the  $\alpha 2$  and/or  $\alpha 3$  subunits. We show that their selective activation by the non-sedative (' $\alpha 1$ -sparing') benzodiazepine-site ligand L-838,417 (ref. 13) is highly effective against inflammatory and neuropathic pain yet devoid of unwanted sedation, motor impairment and tolerance development. L-838,417 not only diminished the nociceptive input to the brain but also reduced the activity of brain areas related to the associative-emotional components of pain, as shown by functional magnetic resonance imaging in rats. These results provide a rational basis for the development of subtype-selective GABAergic drugs for the treatment of chronic pain, which is often refractory to classical analgesics.

More than 40 years ago, the gate control theory of pain<sup>14</sup> proposed that inhibitory neurons in the superficial dorsal horn of the spinal cord control the relay of nociceptive signals (that is, those evoked by painful stimuli) from the periphery to higher areas of the central nervous system. The pivotal role of inhibitory GABAergic and glycinergic neurons in this process has recently been demonstrated in several reports indicating that a loss of inhibitory neurotransmission underlies several forms of chronic pain<sup>3–7</sup>. Despite this knowledge, inhibitory neurotransmitter receptors have rarely been considered as targets for analgesic treatment. In fact, classical benzodiazepines, which are routinely used for their sedative, anxiolytic and anticonvulsant activity, largely lack clear analgesic efficacy in humans when given systemically<sup>15</sup>. To address this obvious discrepancy we investigated the molecular basis of GABAergic pain control in the spinal cord in an integrative approach based on an electrophysiological and behavioural analysis of genetically modified mice and on functional imaging in rats.

We first tested whether benzodiazepines exert antinociceptive effects at the level of the spinal cord by employing the mouse formalin assay, a model of tonic chemically induced pain. When the classical

benzodiazepine diazepam was injected intrathecally into the lumbar spinal canal at doses of 0.01–0.09 mg per kg body weight, an apparent dose-dependent and reversible antinociception was obtained that could be antagonized by systemic treatment with the benzodiazepine antagonist flumazenil (10 mg kg<sup>-1</sup> intraperitoneally (i.p.)) (Supplementary Fig. 1).

We next sought to identify the GABA<sub>A</sub> receptor isoforms responsible for this antinociception. GABA<sub>A</sub> receptors are heteropentameric ion channels composed from a repertoire of up to 19 subunits<sup>16</sup>. Benzodiazepine-sensitive isoforms are characterized by the presence of the  $\gamma 2$  subunit and one of four  $\alpha$  subunits ( $\alpha 1$ ,  $\alpha 2$ ,  $\alpha 3$  or  $\alpha 5$ )<sup>17</sup>. The generation of four lines of GABA<sub>A</sub>-receptor point-mutated knock-in mice ( $\alpha 1$ (H101R),  $\alpha 2$ (H101R),  $\alpha 3$ (H126R) and  $\alpha 5$ (H105R)), in which a conserved histidine residue had been mutated to arginine, rendering the respective subunit insensitive to diazepam, has enabled the attribution of the different actions of diazepam to the individual GABA<sub>A</sub> receptor isoforms<sup>10–12</sup>. It also became possible to attribute the sedative effects of diazepam to GABA<sub>A</sub> receptors containing an  $\alpha 1$  subunit<sup>10</sup> and the anxiolytic effect to those containing an  $\alpha 2$  subunit<sup>11</sup> or—at high receptor occupancy—an  $\alpha 3$  subunit<sup>18</sup>. We then compared the antinociceptive efficacy of intrathecal diazepam (0.09 mg kg<sup>-1</sup>) in wild-type mice with that obtained in the four types of GABA<sub>A</sub>-receptor point-mutated mice in models of inflammatory hyperalgesia induced by subcutaneous injection of zymosan A into one hindpaw and of neuropathic pain evoked by chronic constriction of the left sciatic nerve (chronic constriction injury (CCI) model).

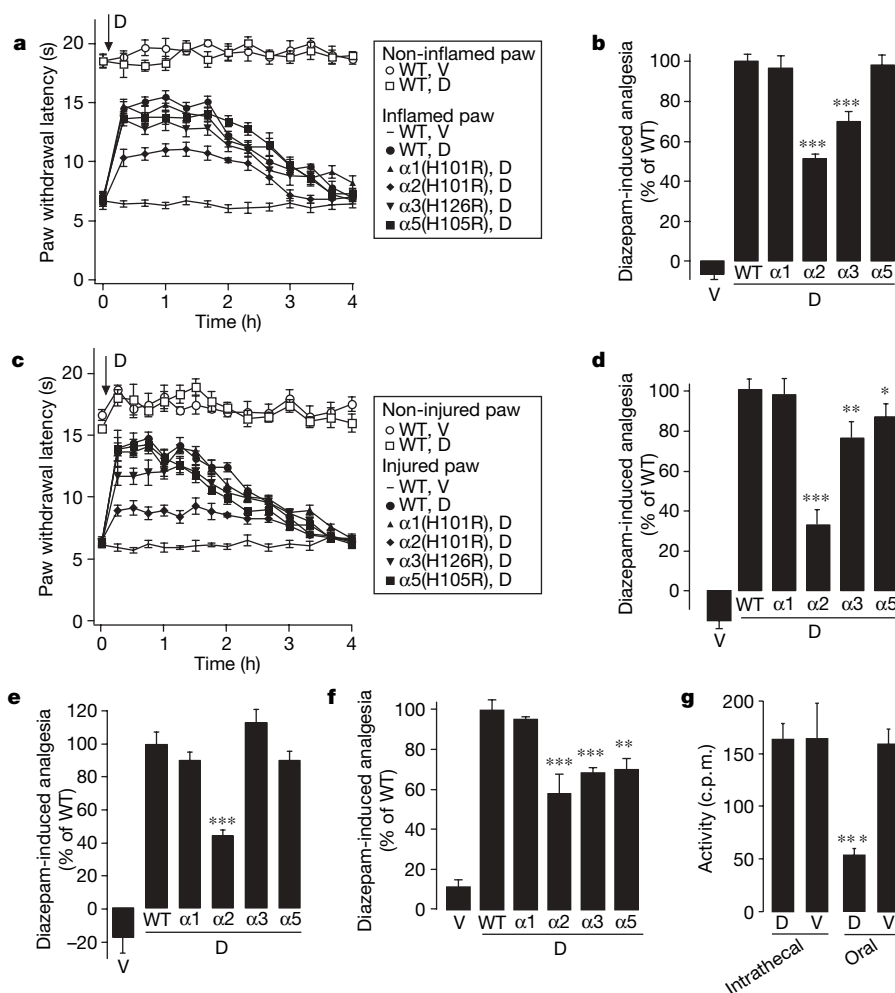
Wild-type mice and all four types of mutant mice developed nearly identical pain sensitization after induction of inflammation or peripheral nerve injury (Fig. 1a, c). In wild-type mice, intrathecal diazepam (0.09 mg kg<sup>-1</sup>) reversibly reduced inflammatory heat hyperalgesia (Fig. 1b), as well as CCI-induced heat hyperalgesia (Fig. 1d), cold allodynia (Fig. 1e) and mechanical sensitization (Fig. 1f) by  $82 \pm 13\%$ ,  $92 \pm 6\%$  and  $79 \pm 9\%$  (means  $\pm$  s.e.m.), respectively. Responses of the non-inflamed or uninjured side were not significantly changed (Fig. 1a, c), indicating that spinal diazepam acted as an anti-hyperalgesic agent rather than as a general analgesic. Almost identical anti-hyperalgesic effects to those in wild-type mice were seen in mice carrying diazepam-insensitive  $\alpha 1$  subunits. By contrast,  $\alpha 2$ (H101R) mice showed a pronounced reduction in diazepam-induced anti-hyperalgesia, which was consistently observed in all pain models tested.  $\alpha 3$ (H126R) and  $\alpha 5$ (H105R) mice showed smaller reductions, which occurred only in a subset of models. Importantly, intrathecal diazepam did not change spontaneous motor activity (Fig. 1g), indicating that the action of diazepam

<sup>1</sup>Institute of Experimental and Clinical Pharmacology and Toxicology, University of Erlangen-Nürnberg, D-91054 Erlangen, Germany. <sup>2</sup>Institute of Pharmacology and Toxicology, University of Zurich, CH-8057 Zurich, Switzerland. <sup>3</sup>Institute of Pharmaceutical Sciences, ETH Zurich, CH-8093 Zurich, Switzerland. <sup>4</sup>Laboratory of Genetic Neuropharmacology, McLean Hospital, Department of Psychiatry, Harvard Medical School, Belmont, Massachusetts 02478, USA. <sup>5</sup>Collegium Helveticum, CH-8092 Zurich, Switzerland. <sup>†</sup>Present addresses: Department of Physiology, University of Bonn, D-53111 Bonn, Germany (S.A.); Department of Physiology, University of Münster, D-48149 Münster, Germany (J.B.).

remained restricted to the spinal level and did not reach supraspinal sites, where sedation would have been induced.

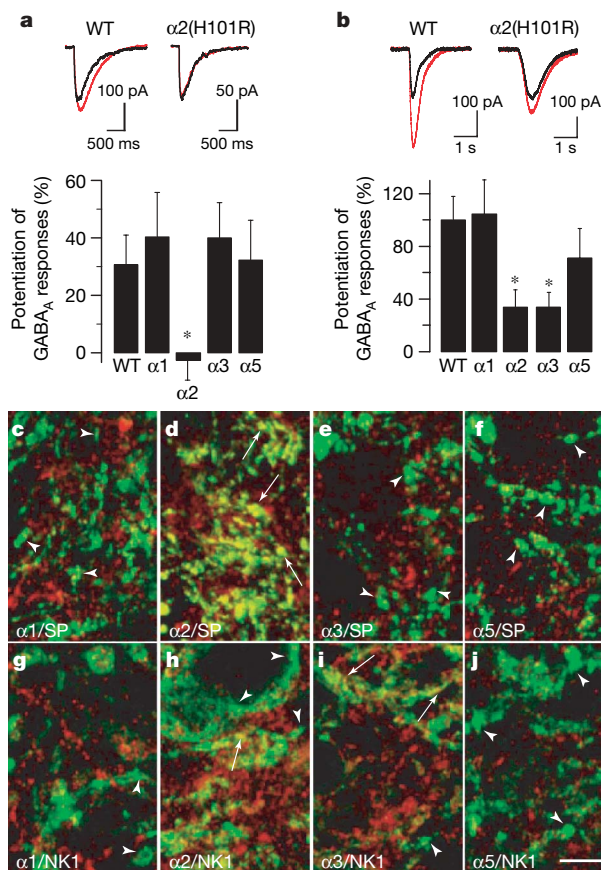
Anti-hyperalgesic effects of spinal diazepam can in principle originate from the facilitation of GABA<sub>A</sub> receptors at different sites. Diazepam might act either on postsynaptic GABA<sub>A</sub> receptors located on intrinsic dorsal horn neurons, thereby increasing postsynaptic inhibition, or on GABA<sub>A</sub> receptors located on the central terminals of primary afferent nerve fibres to increase primary afferent depolarization and presynaptic inhibition<sup>19</sup>. To identify the benzodiazepine-sensitive GABA<sub>A</sub> receptor isoforms expressed at these sites we first employed electrophysiological measurements. GABAergic membrane currents were recorded from superficial dorsal horn neurons in transverse slices of spinal cords and from acutely isolated primary afferent (dorsal root ganglion (DRG)) nociceptive neurons characterized by their sensitivity to capsaicin. In nociceptive DRG neurons obtained from  $\alpha 2$ (H101R) mice, the facilitation of GABAergic membrane currents by diazepam was completely abolished, whereas no significant alteration was found in neurons from  $\alpha 1$ (H101R),  $\alpha 3$ (H126R) and  $\alpha 5$ (H105R) mice (Fig. 2a). Facilitation of GABAergic membrane currents by diazepam in intrinsic superficial dorsal horn (lamina I/II) neurons was significantly decreased in

$\alpha 2$ (H101R) and  $\alpha 3$ (H126R) mice but not in  $\alpha 1$ (H101R) or  $\alpha 5$ (H105R) mice (Fig. 2b). We next employed confocal immunofluorescence microscopy of dorsal horn GABA<sub>A</sub> receptor  $\alpha$  subunits and studied their colocalization with substance P (a marker for primary peptidergic nociceptors) and for neurokinin 1 (NK1) receptors (a marker for intrinsic nociceptive dorsal horn neurons in lamina I). Consistent with our electrophysiological experiments and with previous morphological results in the rat<sup>20</sup> was our observation that  $\alpha 2$  and  $\alpha 3$  were the most abundant diazepam-sensitive GABA<sub>A</sub> receptor  $\alpha$  subunits in the mouse spinal dorsal horn (Supplementary Fig. 2). Co-staining experiments with antibodies against substance P or NK1 receptors (Fig. 2c–j and Supplementary Table 1) revealed that  $\alpha 2$ , but not  $\alpha 1$ ,  $\alpha 3$  or  $\alpha 5$ , were extensively colocalized with substance-P-positive primary afferent terminals in lamina II, whereas colocalization with NK1-receptor-positive lamina I neurons was greatest for the  $\alpha 3$  subunit. Staining for  $\alpha 1$  and  $\alpha 5$  subunits was much less abundant and only occasionally colocalized with either substance P or NK1 receptors. Both sets of experiments indicate that intrinsic dorsal horn neurons express mainly GABA<sub>A</sub> receptor isoforms containing  $\alpha 2$  and  $\alpha 3$  subunits, whereas  $\alpha 2$  is the dominant diazepam-sensitive GABA<sub>A</sub> receptor  $\alpha$  subunit in adult DRG neurons (see also ref. 21).



**Figure 1** | Antinociceptive effects of spinal diazepam in different mouse pain models. **a, b**, Inflammatory pain induced by subcutaneous injection of zymosan A into the left hindpaw in wild-type (WT) mice and GABA<sub>A</sub> receptor point-mutated mice ( $\alpha 1$ (H101R),  $\alpha 2$ (H101R),  $\alpha 3$ (H126R),  $\alpha 5$ (H105R)). **a**, Paw withdrawal latencies (mean  $\pm$  s.e.m.) in response to a defined radiant heat stimulus versus time after administration of intrathecal diazepam (D; 0.09 mg kg<sup>-1</sup>; arrowed) 48 h after injection of zymosan A. V, vehicle. **b**, Percentage diazepam-induced analgesia in the different genotypes. **c, d**, As in **a** and **b**, but for the CCI model of neuropathic pain.

**e, f**, Effects of intrathecal diazepam (0.09 mg kg<sup>-1</sup>) on cold allodynia (**e**) and mechanical sensitivity (**f**) seven days after CCI surgery. Asterisk,  $P \leq 0.05$ ; two asterisks,  $P \leq 0.01$ ; three asterisks,  $P \leq 0.001$  (statistically significant against wild type; ANOVA followed by Bonferroni post-hoc test,  $n = 6$  or 7 mice per group). **g**, Effects of diazepam (0.09 mg kg<sup>-1</sup> intrathecally, or 10 mg kg<sup>-1</sup> orally) on motor activity in the Actifram test (mean  $\pm$  s.e.m.,  $n = 5$  or 6), 10–30 min after intrathecal drug application or 40–80 min after oral drug application. Three asterisks,  $P \leq 0.001$  against vehicle (unpaired  $t$ -test).



The decrease in diazepam-induced anti-hyperalgesia in  $\alpha 2$ (H101R) and  $\alpha 3$ (H126R) mice corresponds well to the presence of these subunits on primary afferent nerve terminals and/or on intrinsic dorsal horn neurons.

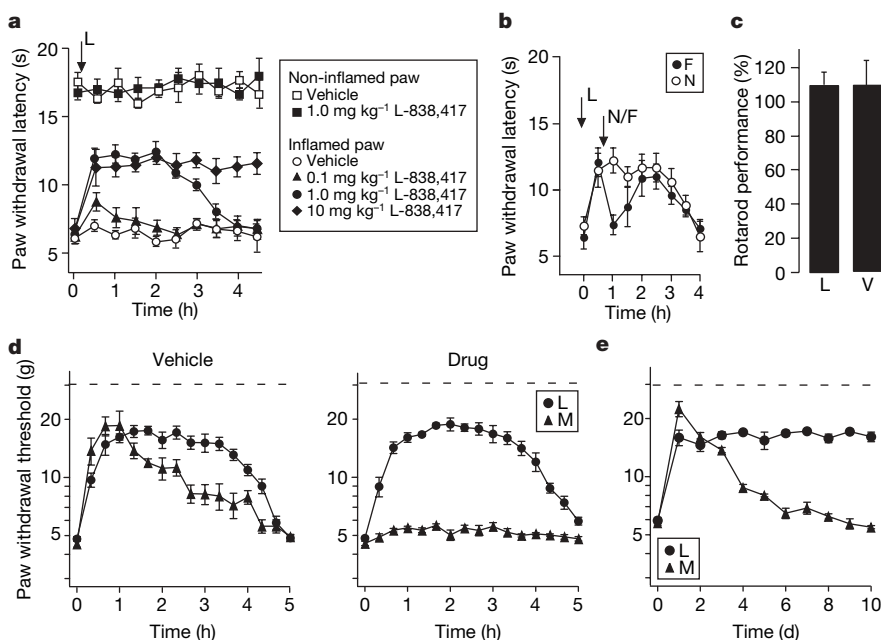
So far, our results indicated that the spinal antinociceptive effect of diazepam is mainly mediated by GABA<sub>A</sub> receptor isoforms containing the  $\alpha 2$  and  $\alpha 3$  subunits, whereas the activation of  $\alpha 1$ -containing GABA<sub>A</sub> receptors is not involved. We therefore tested whether a similar analgesic effect would also be achieved after systemic treatment with subtype-selective benzodiazepine-site agonists, which spare the

**Figure 2 | GABA<sub>A</sub> receptor  $\alpha$  subunits in capsaicin-sensitive primary afferent DRG neurons and in intrinsic dorsal horn neurons.**  
**a, b**, Potentiation of GABAergic membrane currents by diazepam in wild-type (WT) and GABA<sub>A</sub> receptor mutant mice. **a**, DRG neurons. Averaged membrane currents evoked by puff-applied exogenous GABA (1 mM) and percentage potentiation (mean  $\pm$  s.e.m.) by diazepam (1  $\mu$ M,  $n = 5-9$ ). Asterisk,  $P \leq 0.05$  (significant against all other genotypes; ANOVA followed by Fisher's post-hoc test). **b**, Intrinsic superficial dorsal horn neurons (mean  $\pm$  s.e.m.,  $n = 5-10$ ). Asterisk,  $P \leq 0.05$  (significant against wild-type and  $\alpha 1$ (H101R); ANOVA followed by Fisher's post-hoc test). **c-j**, Double immunofluorescence staining showing differential distribution of GABA<sub>A</sub> receptor  $\alpha$  subunits (red) relative to substance P (SP)-positive axons and terminals (green) (**c-f**) or NK1 receptor-positive neurons (**g-j**) in laminae I and II. **c, g**,  $\alpha 1$ ; **d, h**,  $\alpha 2$ ; **e, i**,  $\alpha 3$ ; **f, j**,  $\alpha 5$ . Arrows, double-labelled structures. Arrowheads, single-labelled structures devoid of GABA<sub>A</sub> receptor labelling. Scale bar, 5  $\mu$ m (**c-j**).

$\alpha 1$  subunit, by employing the non-sedative benzodiazepine-site ligand L-838,417, which is an antagonist at the  $\alpha 1$  subunit and a partial agonist at receptors containing  $\alpha 2$ ,  $\alpha 3$  and  $\alpha 5$  subunits<sup>13</sup>.

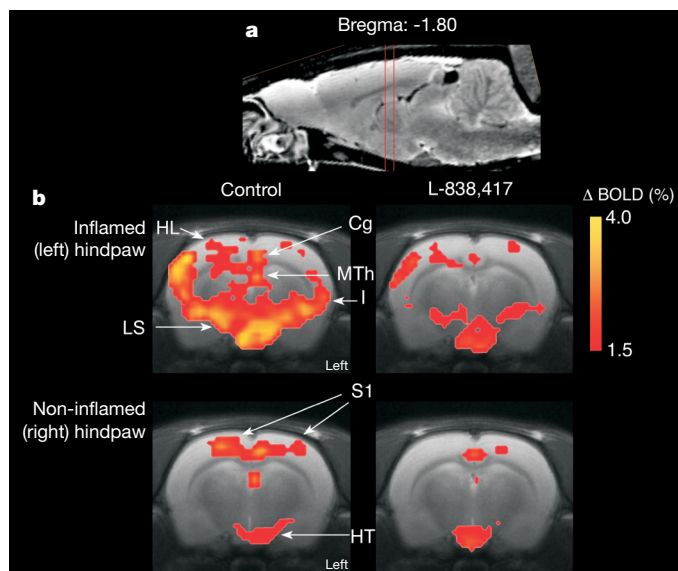
Because L-838,417 possesses poor bioavailability and an extremely short half-life in mice<sup>22</sup>, it was tested in rats. After systemic treatment, L-838,417 produced dose-dependent and reversible anti-hyperalgesia in both the inflammatory and neuropathic pain models (Fig. 3). As expected, its maximum anti-hyperalgesic effect (Fig. 3a) was less than that of intrathecal diazepam, probably because L-838,417 exerts only partial agonistic activity. Anti-hyperalgesia was again completely reversed by flumazenil (10 mg kg<sup>-1</sup> i.p.; Fig. 3b), indicating that it was mediated through the benzodiazepine-binding site of GABA<sub>A</sub> receptors. It was, however, insensitive to the opioid receptor antagonist naloxone (10 mg kg<sup>-1</sup> i.p.), demonstrating that opioidergic pathways were not involved (Fig. 3b). L-838,417 did not impair motor coordination (Fig. 3c). We next investigated the effects of L-838,417 against neuropathic pain and compared its analgesic efficacy and its liability to tolerance development (that is, its loss of analgesic activity) with that of morphine. L-838,417 had a maximum analgesic effect comparable to that of morphine (20 mg kg<sup>-1</sup> i.p.) (Fig. 3d), but unlike morphine it did not lose its efficacy during a chronic (nine-day) treatment period (Fig. 3d, e).

Finally, functional magnetic resonance imaging (fMRI) was used to assess whether L-838,417 would reduce not only nociceptive behaviour but also the representation of pain in the central nervous system. Changes in blood-oxygenation-level-dependent (BOLD) signals were quantified to measure brain activation evoked by noxious



**Figure 3 | Anti-hyperalgesic effects of the non-sedative benzodiazepine site ligand L-838,417 in rats.**

**a, b**, Inflammatory hyperalgesia induced by subcutaneous injection of zymosan A (1 mg) into one hindpaw. **a**, Effects of administration of L-838,417 (arrowed) on thermal hyperalgesia 6 h after injection of zymosan A ( $n = 4-6$  rats). **b**, Effects of the benzodiazepine site antagonist flumazenil (F; 10 mg kg<sup>-1</sup> i.p.) and the opioid receptor antagonist naloxone (N; 10 mg kg<sup>-1</sup> i.p.) on antinociception induced by administration of L-838,417 (L; 1 mg kg<sup>-1</sup> orally).  $n = 3$  rats per group. **c**, Effects of L-838,417 (1 mg kg<sup>-1</sup> orally) on motor control, shown as percentages of pre-drug rotarod performance ( $n = 8$  rats per group). **d, e**, Neuropathic pain induced by CCI surgery. **d**, Anti-hyperalgesia by L-838,417 and morphine after chronic treatment (once-daily i.p. injections) for 9 days with either drug (right) or vehicle (left), 16 days after CCI surgery. Dashed lines, thresholds before CCI surgery. **e**, Analgesic efficacy of L-838,417 (L, 1 mg kg<sup>-1</sup>) and morphine (M, 20 mg kg<sup>-1</sup>) versus treatment duration.  $n = 6$  rats per group. For a comparison with the anti-hyperalgesic activity of intrathecal diazepam in rats see Supplementary Fig. 3. All data are means  $\pm$  s.e.m.



**Figure 4 | Effects of L-838,417 (1 mg kg<sup>-1</sup> i.p.) on the supraspinal representation of pain.** **a**, Anatomical slice indicating the position of the functional images. **b**, False-colour images of changes in BOLD signals evoked by stimulation of the left (inflamed) or right (non-inflamed) hindpaw with noxious heat. Images represent group maps across 12 rats averaged from 8 (pre-drug) and 16 (post-drug) stimulations. Experiments started 6 h after subcutaneous zymosan A injection into the left hindpaw. MTh, medial thalamus; S1, primary somatosensory cortex; Cg, cingulate cortex; I, insular cortex; LS, limbic system (including amygdala, entorhinal cortex and hippocampus); HT, hypothalamus; HL, representation of hindlimb in S1. Left, left hemisphere.

heat. Stimulation of the inflamed left or the non-inflamed right hindpaw led to reliable, often bilateral, activation of several brain regions involved in pain processing (Fig. 4). Significantly more brain volume was activated and stronger activation was seen on stimulation of the inflamed paw. L-838,417 (1 mg kg<sup>-1</sup> i.p.) decreased brain activation induced by noxious heat after stimulation of the inflamed paw. For a quantitative assessment of its analgesic effects, we integrated the stimulus-correlated change in the BOLD signal ( $F$ ) over all significantly activated voxels of each region of interest and calculated  $\Delta F/F$  as  $(F_{\text{post}} - F_{\text{pre}})/F_{\text{pre}}$ , the relative decrease in  $F$  after injection of L-838,417 (Table 1) or vehicle (Supplementary Table 2). Here we focused on brain areas that reflected either the sensory and discriminative component of pain (the medial thalamus and contralateral primary sensory cortex) or its emotional dimension (limbic system and frontal association cortex)<sup>23,24</sup>. After stimulation of the inflamed paw, a pronounced and statistically significant reduction in BOLD signal changes was observed in most brain regions analysed. Smaller changes in brain activation were found when the non-inflamed paw was stimulated, and only negligible effects were seen after innocuous thermal stimulation (Table 1). These results indicate that systemically administered L-838,417 does indeed act as an anti-hyperalgesic agent and reduces BOLD signals in brain areas

related to both the sensory and the emotional associative components of pain.

Considerable evidence indicates that a facilitation of GABAergic inhibition can be pro-nociceptive at supraspinal sites, for example the rostral agranular insular cortex<sup>25</sup> or in the periaqueductal grey<sup>26</sup>, by reducing the activity of descending antinociceptive neurons. At these sites most GABA<sub>A</sub> receptors apparently contain the  $\alpha 1$  subunit<sup>27</sup>. Therefore, not only would sparing the  $\alpha 1$  subunit avoid unwanted sedation, it would also increase analgesic efficacy. Aside from sedation and tolerance development, addictive properties are of major concern in the development of analgesics. Available evidence indicates that subtype-selective benzodiazepine-site ligands should exhibit at most only modest addictive properties<sup>28</sup> and should not lead to tolerance development<sup>29</sup>. Finally, previous studies have shown that in neuropathic pain after injury to peripheral nerves, GABAergic inhibition can not only be diminished but it can even turn into excitation<sup>6,7</sup>. Our results suggest that sufficient inhibition remains to permit a spinal analgesic effect of drugs that increase GABAergic neurotransmission. Because glycine and GABA are released together at many inhibitory synapses in the dorsal horn<sup>30</sup>, a facilitation of GABAergic transmission should also be able to compensate for a selective decrease in glycinergic inhibition<sup>3</sup>. Thus, we have not only identified the GABA<sub>A</sub> receptors containing the  $\alpha 2$  and  $\alpha 3$  subunits as critical components of spinal pain control, but also demonstrated that  $\alpha 1$ -sparing benzodiazepine-site ligands, which are already in development as anxiolytic (non-sedative) agents, might constitute a class of analgesics suitable for the treatment of chronic pain syndromes.

## METHODS SUMMARY

Wild-type mice and GABA<sub>A</sub> receptor mutant mice ( $\alpha 1$ (H101R),  $\alpha 2$ (H101R),  $\alpha 3$ (H126R) and  $\alpha 5$ (H105R))<sup>10–12</sup> were maintained on a 129X1/SvJ background. Behavioural experiments were performed on adult mice and Wistar rats. Chemically induced pain was assessed in the formalin test, in which flinches of the injected paw were counted for 60 min after subcutaneous injection of 5% formalin into one hindpaw. Inflammatory pain was induced by subcutaneous injection of zymosan A (0.06 mg in mice; 1 mg in rats) into one hindpaw. Neuropathic pain was studied after chronic constriction of the left sciatic nerve. Heat hyperalgesia was assessed by measuring paw withdrawal latencies on exposure to defined radiant heat. Cold allodynia was measured as the time spent lifting, shaking or licking the paw (seconds per minute) after the application of acetone onto the affected paw. Mechanical sensitivity was assessed with electronic von Frey filaments. Locomotor activity was measured by using microprocessor-controlled activity cages, and motor function was assessed in the rotarod test. In all behavioural tests, the observer was blinded to the genotype or to the drug treatment.

GABAergic membrane currents were recorded from acutely dissociated capsaicin-sensitive DRG neurons (segments L4–L6) and from superficial dorsal horn neurons (layers I and II) in transverse slices of lumbar spinal cord, both obtained from 14–24-day-old mice.

fMRI experiments were performed on adult male Wistar rats slightly anaesthetized with 1–2% isoflurane, using a Bruker 4.7-T Biospec scanner. Heat stimulation was performed by applying temperature ramps to 52 °C and to 42 °C (noxious and innocuous stimulation, respectively) through Peltier elements tightly attached to the hindpaws.

The localization of GABA<sub>A</sub> receptor  $\alpha$  subunits on primary afferent nerve terminals and on intrinsic dorsal horn neurons was determined by double

**Table 1 | Changes in heat-induced brain activation by L-838,417 measured by rat fMRI**

Area	Stimulation of inflamed paw with noxious heat		Stimulation of non-inflamed paw with noxious heat		Stimulation of non-inflamed paw with innocuous temperature	
	$\Delta F/F$ ; incidence*	P†	$\Delta F/F$ ; incidence*	P†	$\Delta F/F$ ; incidence*	P†
MTh	$-0.35 \pm 0.07$ ; 10/12	0.014	$-0.29 \pm 0.09$ ; 10/12	0.069	$-0.10 \pm 0.07$ ; 6/12	0.302
S1c	$-0.29 \pm 0.07$ ; 12/12	0.028	$-0.07 \pm 0.09$ ; 12/12	0.383	$-0.18 \pm 0.25$ ; 11/12	0.228
Cg	$-0.37 \pm 0.07$ ; 11/12	0.034	$-0.26 \pm 0.08$ ; 12/12	0.078	$-0.07 \pm 0.08$ ; 10/12	0.152
FAC	$-0.55 \pm 0.05$ ; 12/12	0.007	$-0.30 \pm 0.08$ ; 12/12	0.179	$-0.09 \pm 0.08$ ; 6/12	0.253
LS	$-0.36 \pm 0.05$ ; 11/12	0.012	$-0.06 \pm 0.07$ ; 12/12	0.580	$-0.04 \pm 0.07$ ; 11/12	0.413

Where errors are shown, results are means  $\pm$  s.e.m. MTh, medial thalamus; S1c, contralateral primary somatosensory cortex; Cg, cingulate cortex; FAC, frontal association cortex; LS, limbic system (including amygdala, entorhinal cortex and hippocampus).

\* Number of rats in which a significant noxious heat-induced activation of the respective area occurred/total number of rats studied.

† Significance versus pre-drug (paired Student *t*-test).

immunofluorescence staining on sections from perfusion-fixed adult mice<sup>27</sup>. Confocal images were processed with Imaris (Bitplane).

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Received 11 October; accepted 19 November 2007.**

- Sandkühler, J. Learning and memory in pain pathways. *Pain* **88**, 113–118 (2000).
- Woolf, C. J. & Salter, M. W. Neuronal plasticity: increasing the gain in pain. *Science* **288**, 1765–1769 (2000).
- Ahmadi, S., Lippross, S., Neuhuber, W. L. & Zeilhofer, H. U. PGE<sub>2</sub> selectively blocks inhibitory glycinergic neurotransmission onto rat superficial dorsal horn neurons. *Nature Neurosci.* **5**, 34–40 (2002).
- Harvey, R. J. *et al.* GlyR $\alpha$ 3: an essential target for spinal PGE<sub>2</sub>-mediated inflammatory pain sensitization. *Science* **304**, 884–887 (2004).
- Moore, K. A. *et al.* Partial peripheral nerve injury promotes a selective loss of GABAergic inhibition in the superficial dorsal horn of the spinal cord. *J. Neurosci.* **22**, 6724–6731 (2002).
- Coull, J. A. *et al.* Trans-synaptic shift in anion gradient in spinal lamina I neurons as a mechanism of neuropathic pain. *Nature* **424**, 938–942 (2003).
- Coull, J. A. *et al.* BDNF from microglia causes the shift in neuronal anion gradient underlying neuropathic pain. *Nature* **438**, 1017–1021 (2005).
- Scholz, J. *et al.* Blocking caspase activity prevents transsynaptic neuronal apoptosis and the loss of inhibition in lamina II of the dorsal horn after peripheral nerve injury. *J. Neurosci.* **25**, 7317–7323 (2005).
- Malan, T. P., Mata, H. P. & Porreca, F. Spinal GABA<sub>A</sub> and GABA<sub>B</sub> receptor pharmacology in a rat model of neuropathic pain. *Anesthesiology* **96**, 1161–1167 (2002).
- Rudolph, U. *et al.* Benzodiazepine actions mediated by specific  $\gamma$ -aminobutyric acid<sub>A</sub> receptor subtypes. *Nature* **401**, 796–800 (1999).
- Löw, K. *et al.* Molecular and neuronal substrate for the selective attenuation of anxiety. *Science* **290**, 131–134 (2000).
- Crestani, F. *et al.* Trace fear conditioning involves hippocampal  $\alpha$ 5 GABA<sub>A</sub> receptors. *Proc. Natl Acad. Sci. USA* **99**, 8980–8985 (2002).
- McKernan, R. M. *et al.* Sedative but not anxiolytic properties of benzodiazepines are mediated by the GABA<sub>A</sub> receptor  $\alpha$ 1 subtype. *Nature Neurosci.* **3**, 587–592 (2000).
- Melzack, R. & Wall, P. D. Pain mechanisms: a new theory. *Science* **150**, 971–979 (1965).
- Enna, S. J. & McCarson, K. E. The role of GABA in the mediation and perception of pain. *Adv. Pharmacol.* **54**, 1–27 (2006).
- Barnard, E. A. *et al.* International Union of Pharmacology. XV. Subtypes of  $\gamma$ -aminobutyric acid A receptors: classification on the basis of subunit structure and receptor function. *Pharmacol. Rev.* **50**, 291–313 (1998).
- Wieland, H. A., Lüddens, H. & Seeburg, P. H. A single histidine in GABA<sub>A</sub> receptors is essential for benzodiazepine agonist binding. *J. Biol. Chem.* **267**, 1426–1429 (1992).
- Dias, R. *et al.* Evidence for a significant role of  $\alpha$ 3-containing GABA<sub>A</sub> receptors in mediating the anxiolytic effects of benzodiazepines. *J. Neurosci.* **25**, 10682–10688 (2005).
- Rudomin, P. & Schmidt, R. F. Presynaptic inhibition in the vertebrate spinal cord revisited. *Exp. Brain Res.* **129**, 1–37 (1999).
- Bohlhalter, S., Weinmann, O., Möhler, H. & Fritschy, J. M. Laminar compartmentalization of GABA<sub>A</sub>-receptor subtypes in the spinal cord: an immunohistochemical study. *J. Neurosci.* **16**, 283–297 (1996).
- Ma, W., Saunders, P. A., Somogyi, R., Poulter, M. O. & Barker, J. L. Ontogeny of GABA<sub>A</sub> receptor subunit mRNAs in rat spinal cord and dorsal root ganglia. *J. Comp. Neurol.* **338**, 337–359 (1993).
- Scott-Stevens, P., Atack, J. R., Sohal, B. & Worboys, P. Rodent pharmacokinetics and receptor occupancy of the GABA<sub>A</sub> receptor subtype selective benzodiazepine site ligand L-838417. *Biopharm. Drug Dispos.* **26**, 13–20 (2005).
- Brooks, J. & Tracey, I. From nociception to pain perception: imaging the spinal and supraspinal pathways. *J. Anat.* **207**, 19–33 (2005).
- Bushnell, M. C. & Apkarian, A. V. in *Wall and Melzack's Textbook of Pain* (ed. McMahon, S. B. & Koltzenburg, M.) 107–124 (Elsevier Churchill Livingstone, London, 2006).
- Jasmin, L., Rabkin, S. D., Granato, A., Boudah, A. & Ohara, P. T. Analgesia and hyperalgesia from GABA-mediated modulation of the cerebral cortex. *Nature* **424**, 316–320 (2003).
- Harris, J. A. & Westbrook, R. F. Effects of benzodiazepine microinjection into the amygdala or periaqueductal gray on the expression of conditioned fear and hypoaesthesia in rats. *Behav. Neurosci.* **109**, 295–304 (1995).
- Fritschy, J. M. & Möhler, H. GABA<sub>A</sub>-receptor heterogeneity in the adult rat brain: differential regional and cellular distribution of seven major subunits. *J. Comp. Neurol.* **359**, 154–194 (1995).
- Ator, N. A. Contributions of GABA<sub>A</sub> receptor subtype selectivity to abuse liability and dependence potential of pharmacological treatments for anxiety and sleep disorders. *CNS Spectr.* **10**, 31–39 (2005).
- van Rijnsoever, C. *et al.* Requirement of  $\alpha$ 5-GABA<sub>A</sub> receptors for the development of tolerance to the sedative action of diazepam in mice. *J. Neurosci.* **24**, 6785–6790 (2004).
- Keller, A. F., Coull, J. A., Chery, N., Poisbeau, P. & de Koninck, Y. Region-specific developmental specialization of GABA-glycine cosynapses in laminae I–II of the rat spinal dorsal horn. *J. Neurosci.* **21**, 7871–7880 (2001).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank M. Rudin for critical reading of the manuscript, and R. Keist, I. Camenisch, B. Layh, S. Gabriel, C. Sidler and S. John for technical assistance. This work was supported by grants from the Deutsche Forschungsgemeinschaft to H.U.Z. and A.H., by the Bundesministerium für Bildung und Forschung (migraine and BCCN) to A.H., by grants from the Schweizerischer Nationalfonds to J.M.F., H.M., U.R. and H.U.Z., the NCCR Neural Plasticity and Repair, and by the Doerenkamp Foundation for Innovations in Animal and Consumer Protection to K.B.

**Author Contributions** J.K., R.W., K.H., H.R. and U.B.Z. conducted the behavioural experiments. S.A. and J.B. made the electrophysiological recordings and analyses. M.S., A.H. and K.B. performed the fMRI study. J.M.F. made the morphological analyses. U.R. and H.M. provided the four lines of genetically modified mice. H.M. suggested experiments with L-838,417. H.U.Z. initiated the research, analysed behavioural and electrophysiological data and wrote the manuscript. All authors made comments on the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence and requests for materials should be addressed to H.U.Z. ([zeilhofer@pharma.uzh.ch](mailto:zeilhofer@pharma.uzh.ch)).

## METHODS

**Mice and rats.** Behavioural experiments were performed in male and female 7–12-week-old mice or in male 7–12-week-old Wistar rats. Permission for the animal experiments was obtained from the Regierung von Mittelfranken (ref. no. 612-2531.31-17/03) and from the Veterinärämter des Kantons Zürich (ref. no. 121/2006 and 34/2007).

**Drugs.** For intrathecal injection in mice, diazepam was dissolved in 10% dimethyl sulfoxide (DMSO), 90% artificial cerebrospinal fluid (ACSF) (vehicle). Total intrathecal injection volume was 5  $\mu$ l (for details of the injection procedure see ref. 31). Up to a concentration of 20%, intrathecal DMSO had no effect on pain behaviour in mice. For i.p. injection, diazepam was dissolved in 0.3% Tween 80, 99.7% ACSF. Morphine was dissolved in ACSF. L-838,417 synthesized by Anawa was suspended in 0.5% methylcellulose and 0.9% NaCl and was applied to rats either orally or i.p. in a total volume of 200  $\mu$ l. Flumazenil (10 mg kg<sup>-1</sup>) and naloxone (10 mg kg<sup>-1</sup>) were dissolved in DMSO (1%) and injected i.p. in a total volume of 200  $\mu$ l.

**Formalin test.** Formalin (5%, 20  $\mu$ l) was injected subcutaneously into the dorsal surface of the left hindpaw<sup>32</sup>. Flinches of the injected paw were counted for 60 min starting immediately after formalin injection. Intrathecal drugs (diazepam or vehicle) were injected 10 min before formalin injection. Flumazenil (10 mg kg<sup>-1</sup>) was injected i.p. 30 min before formalin injection.

**Inflammatory pain.** Inflammatory pain was assessed in the zymosan A model<sup>33</sup>. In mice, 0.06 mg of zymosan A suspended in 20  $\mu$ l of 0.9% NaCl was injected subcutaneously into the plantar side of the left hindpaw. The model was also used in rats, but 1 mg of zymosan A was used. Heat hyperalgesia was assessed 24 h and 6 h after induction of inflammation in mice and rats, respectively.

**Neuropathic pain.** Diazepam, L-838,417 and morphine were analysed in the CCI model<sup>34</sup> in 7–12-week-old mice or rats. Unilateral constriction injury of the left sciatic nerve just proximal to the trifurcation was performed with three loose ligatures. In sham-operated animals the sciatic nerve was exposed and the connective tissue was freed, but no ligatures were applied. In these sham-operated animals only a minor and transient hyperalgesia occurred. Heat hyperalgesia, cold allodynia and mechanical sensitization were assessed 7–9 days after surgery.

**Heat hyperalgesia.** Paw withdrawal latencies on exposure to a defined radiant heat stimulus were measured with a commercially available apparatus (Plantar Test; Ugo Basile). Four or five measurements were taken in each animal for every time point. Measurements of paw withdrawal latencies of the inflamed or injured paw and of the contralateral paw were made alternately.

**Cold allodynia.** The time spent lifting, shaking or licking the paw (seconds per minute) was measured for 5 min after application of a drop of acetone onto the affected paw.

**Mechanical sensitization.** Mechanical sensitivity was assessed with electronic von Frey filaments (IITC). Triple measurements of paw withdrawal thresholds (g) were made for each time point and animal.

**Locomotor activity.** Locomotor activity was tested with a commercially available microprocessor-controlled activity cage (Actiframe; Gerb Elektronik). Mice were placed in the apparatus 15 min before testing. Motor activity was measured 10–30 min and 40–80 min after intrathecal and oral drug application, respectively.

**Motor impairment.** A possible impairment of motor function was assessed with the rotarod test<sup>35</sup>. Rats were trained on day zero and the maximum speed tolerated for at least 2 min was determined for each rat. On the following day, rotarod performance was determined again 30 min after treatment with L-838,417 or vehicle (administered orally).

**Electrophysiology.** DRGs (from segments L4–L6) were removed from 14–24-day-old mice, dissociated and plated on poly-(L-lysine)-coated cover slips (for details see ref. 36). GABA-induced currents were recorded from capsaicin-sensitive DRG neurons 3–30 h after plating. Transverse slices (250  $\mu$ m thick) of the lumbar spinal cord were prepared from 14–24-day-old mice. GABAergic membrane currents were recorded from superficial dorsal horn neurons (laminae I and II) as described previously<sup>3</sup>. In both preparations, GABA (1 mM) was applied by short (10 ms) puffer applications to the soma of the recorded neuron at a frequency of 0.07 Hz. Diazepam (1  $\mu$ M) was applied by means of bath perfusion. All recordings were made in the presence of the GABA<sub>B</sub> receptor antagonist CGP-55,845 (200  $\mu$ M).

**Immunofluorescence.** The localization of GABA<sub>A</sub> receptor  $\alpha$  subunits on primary afferent nerve terminals and intrinsic dorsal horn neurons was determined by double immunofluorescence staining on sections from perfusion-fixed adult mice<sup>27</sup>. Antibodies were home-made subunit-specific antisera<sup>27</sup> and commercial antibodies against substance P (T1609; Bachem) and NK1 (S8305; Sigma). Sections processed for double immunofluorescence were digitalized by confocal laser scanning microscopy (resolution 90 nm per pixel; two or three images per animal;  $n = 3$  mice) and images were processed with Imaris (Bitplane). Double-labelled objects (image profiles) in single confocal sections were identified by a

segmentation algorithm (minimal size 0.2  $\mu$ m<sup>2</sup>; minimum intensity 50–90 on a 256-grey-level scale). The numbers of single-labelled and double-labelled profiles were calculated. All values are expressed as percentages of double-labelled profiles relative to the marker indicated.

**fMRI methods.** fMRI experiments were performed in male Wistar rats weighing 350–400 g. During the measurements, rats were slightly anaesthetized with isoflurane (1–2%) to maintain a respiration rate of about 60 c.p.m. and constant blood pCO<sub>2</sub> levels. Measurements were made with a Bruker 4.7-T Biospec scanner with a free bore of 40 cm, equipped with an actively radiofrequency-decoupled coil system. A whole-body birdcage resonator enabled homogenous excitation, and a 3-cm quadrature surface coil, which served as a receiver, was located directly above the head of the animal to maximize the signal-to-noise ratio. Constant positioning of the rat's head within the scanner was verified by rapid acquisition of magnetic resonance images at 200-ms intervals. A functional series of 1,470 sets (4 s each, total of 96 min) of 22 axial images (slice thickness 1 mm, field of view 25  $\times$  25 mm<sup>2</sup>, 5.20 to –14.60 mm from the bregma<sup>37</sup>) were acquired with the echo planar imaging technique (EPI: matrix 64  $\times$  64, TR = 4,000 ms, TE<sub>eff</sub> = 23.4 ms, two acquisitions). Anatomical scans with a high spatial resolution were obtained with RARE<sup>38</sup> (slice thickness 1 mm, field of view 25  $\times$  25 mm<sup>2</sup>, matrix 256  $\times$  256, TR = 400 ms, TE = 18 ms, NEX = 8).

Noxious heat stimulation was performed by applying temperature ramps (34–52 °C (noxious stimulation) or 34–42 °C (innocuous stimulation) with 15-s rise and fall times and a 5-s plateau phase) through two Peltier elements tightly attached to both hindpaws (in awake rats this stimulation method yielded paw withdrawal latencies similar to those obtained in the behavioural tests with radiant heat). Thermal stimuli were applied to the left and right hindpaw alternately at 2-min intervals. After 32 min of recording, L-838,417 (1 mg kg<sup>-1</sup>) or vehicle was injected through an i.p. catheter without changing the position of the animal in the scanner. After drug injection, recording was continued for 64 min with the same stimulation method.

Data were analysed with Brainvoyager QX after appropriate preprocessing (motion correction, mean intensity adjustment, spatial smoothing 0.6 mm full-width at half-maximum, temporal gaussian smoothing 12 s, and temporal high-pass filtering of nine cycles) with a General Linear Modelling approach with four predictors: inflamed (left)/non-inflamed (right) paw before and after drug injection and Bonferroni correction.  $z$ -score maps of the individual rats were group analysed with custom-made analysis software (MagnAn<sup>39</sup> running under IDL). Anatomical and functional images were transferred into the register by an affine transformation scheme with only six degrees of freedom derived from the individual brain masks. The registered anatomical data and  $z$ -score maps were averaged over all animals. Contrast-specific mean  $z$ -score maps were calculated using a threshold of 3.0. Significantly activated voxels were labelled automatically with a digital standard rat brain atlas<sup>37</sup>. For each rat, brain structure and stimulation condition, we then first calculated the activation intensity as the stimulus-induced relative change in the BOLD signal ( $F$ ). To quantify the effect of L-838,417 on the stimulus-induced BOLD signal changes we calculated  $\Delta F/F$  as  $(F_{\text{post}} - F_{\text{pre}})/F_{\text{pre}}$ , where  $F_{\text{post}}$  is the value of  $F$  after drug treatment and  $F_{\text{pre}}$  is the value before drug treatment. Statistical analysis was performed with the paired Student  $t$ -test. False-colour images of stimulus-induced changes in BOLD signals were obtained by mapping the calculated mean BOLD signal change of each voxel onto all significantly activated voxels. Note that the different colours in Fig. 4 encode  $F$  (signal amplitude), not statistical coefficients.

31. Depner, U. B., Reinscheid, R. K., Takeshima, H., Brune, K. & Zeilhofer, H. U. Normal sensitivity to acute pain, but increased inflammatory hyperalgesia in mice lacking the nociceptin precursor polypeptide or the nociceptin receptor. *Eur. J. Neurosci.* **17**, 2381–2387 (2003).
32. Dubuisson, D. & Dennis, S. G. The formalin test: a quantitative study of the analgesic effects of morphine, meperidine, and brain stem stimulation in rats and cats. *Pain* **4**, 161–174 (1977).
33. Hargreaves, K., Dubner, R., Brown, F., Flores, C. & Joris, J. A new and sensitive method for measuring thermal nociception in cutaneous hyperalgesia. *Pain* **32**, 77–88 (1988).
34. Bennett, G. J. & Xie, Y. K. A peripheral mononeuropathy in rat that produces disorders of pain sensation like those seen in man. *Pain* **33**, 87–107 (1988).
35. Bonetti, E. P. et al. Ro 15–4513: partial inverse agonism at the BZR and interaction with ethanol. *Pharmacol. Biochem. Behav.* **31**, 733–749 (1988).
36. Zeilhofer, H. U., Kress, M. & Swandulla, D. Fractional Ca<sup>2+</sup> currents through capsaicin- and proton-activated ion channels in rat dorsal root ganglion neurones. *J. Physiol. (Lond.)* **503**, 67–78 (1997).
37. Paxinos, G. & Watson, C. *The Rat Brain in Stereotaxic Coordinates* 4th edn (Academic, San Diego, 1998).
38. Hennig, J., Nauerth, A. & Friedburg, H. RARE imaging: a fast imaging method for clinical MR. *Magn. Reson. Med.* **3**, 823–833 (1986).
39. Hess, A., Sergejeva, M., Budinsky, L., Zeilhofer, H. U. & Brune, K. Imaging of hyperalgesia in rats by functional MRI. *Eur. J. Pain* **11**, 109–119 (2007).

# Identification of *RPS14* as a 5q<sup>−</sup> syndrome gene by RNA interference screen

Benjamin L. Ebert<sup>1,2,3</sup>, Jennifer Pretz<sup>1</sup>, Jocelyn Bosco<sup>1</sup>, Cindy Y. Chang<sup>1</sup>, Pablo Tamayo<sup>1</sup>, Naomi Galili<sup>4</sup>, Azra Raza<sup>4</sup>, David E. Root<sup>1</sup>, Eyal Attar<sup>5</sup>, Steven R. Ellis<sup>6</sup> & Todd R. Golub<sup>1,2,7</sup>

Somatic chromosomal deletions in cancer are thought to indicate the location of tumour suppressor genes, by which a complete loss of gene function occurs through biallelic deletion, point mutation or epigenetic silencing, thus fulfilling Knudson's two-hit hypothesis<sup>1</sup>. In many recurrent deletions, however, such biallelic inactivation has not been found. One prominent example is the 5q<sup>−</sup> syndrome, a subtype of myelodysplastic syndrome characterized by a defect in erythroid differentiation<sup>2</sup>. Here we describe an RNA-mediated interference (RNAi)-based approach to discovery of the 5q<sup>−</sup> disease gene. We found that partial loss of function of the ribosomal subunit protein *RPS14* phenocopies the disease in normal haematopoietic progenitor cells, and also that forced expression of *RPS14* rescues the disease phenotype in patient-derived bone marrow cells. In addition, we identified a block in the processing of pre-ribosomal RNA in *RPS14*-deficient cells that is functionally equivalent to the defect in Diamond–Blackfan anaemia, linking the molecular pathophysiology of the 5q<sup>−</sup> syndrome to a congenital syndrome causing bone marrow failure. These results indicate that the 5q<sup>−</sup> syndrome is caused by a defect in ribosomal protein function and suggest that RNAi screening is an effective strategy for identifying causal haploinsufficiency disease genes.

The 5q<sup>−</sup> syndrome was reported in 1974 as the first chromosomal deletion in cancer associated with a distinct clinical phenotype<sup>2</sup>. Patients have a severe macrocytic anaemia, normal or elevated platelet counts, normal or reduced neutrophil counts, erythroid hypoplasia in the bone marrow, and hypolobated micromegakaryocytes<sup>3</sup>. These patients also have a propensity to progress to acute myeloid leukaemia (AML), although more slowly than other forms of myelodysplastic syndrome (MDS)<sup>4</sup>. A main cause of morbidity and mortality for these patients is the erythroid defect, which often requires continuing transfusions of red blood cells resulting in iron overload and subsequent organ dysfunction<sup>4</sup>. The 5q<sup>−</sup> syndrome is also unique because this subtype of MDS shows a remarkable response to treatment with the thalidomide analogue lenalidomide, although the mechanism of action of lenalidomide remains unknown<sup>5</sup>.

Over the past 30 years, physical mapping methods have been used to narrow the region of recurrent somatic deletion on 5q to a 1.5-megabase common deleted region (CDR) containing 40 genes<sup>6</sup>. No patients with the 5q<sup>−</sup> syndrome have been reported to have biallelic deletions within the CDR, and no point mutations have been reported in the remaining allele of any of the 40 genes in the region. This observation led us to speculate that the 5q<sup>−</sup> syndrome may be caused by haploinsufficiency, suggesting that an alternative approach would be required to identify the gene responsible. We therefore examined whether the principal hallmarks of the disease

(an erythroid maturation block with preservation of megakaryocyte differentiation) could be recapitulated experimentally with short hairpin RNAs (shRNAs) targeting each of the genes within the CDR.

We designed multiple lentivirally expressed shRNAs for each of the candidate genes, to control for possible off-target effects of any individual shRNA. The shRNAs were introduced into normal CD34<sup>+</sup> human haematopoietic progenitor cells, and the cells were induced to differentiate for 10 days along the erythroid and megakaryocytic lineages. The effect of each shRNA was assessed by fluorescence-activated cell sorting (FACS) analysis with the use of erythroid-specific and megakaryocyte-specific cell surface markers. The shRNAs targeting one gene, *RPS14*, recapitulated the 5q<sup>−</sup> syndrome phenotype: a severe decrease in the production of erythroid cells with relative preservation of megakaryocytic cells (Fig. 1). Furthermore, using the sequential expression of CD71 and glycophorin A during erythroid differentiation (Supplementary Fig. 1), we found that shRNAs targeting *RPS14* blocked the production of terminally differentiated erythroid cells, which is also consistent with the 5q<sup>−</sup> syndrome disease phenotype (Supplementary Fig. 2). In a statistical analysis that grouped all shRNAs targeting each gene into a single set, *RPS14* was the only gene that significantly altered differentiation (Supplementary Fig. 3). On the basis of these results, we focused our attention on *RPS14* as a candidate disease gene.

We first confirmed that all five *RPS14* shRNAs that scored in the screen in fact knocked down *RPS14* expression, and that the level of protein expression was of the order of half of the luciferase control cells, which is consistent with a model of *RPS14* haploinsufficiency (Fig. 2a). Each of the *RPS14* shRNAs decreased erythroid differentiation relative to megakaryocytic differentiation (Fig. 2b) and also caused a mild defect in erythroid versus myeloid differentiation (Fig. 2c), precisely as seen in patients with the clinical syndrome. *RPS14* knockdown also caused an increase in the ratio of immature-to-mature erythroid cells (Fig. 2d), as well as increased apoptosis of differentiating erythroid cells (Fig. 2e), which is consistent with the well-described apoptotic phenotype of MDS<sup>7</sup>. Given the possibility that multiple genes in the CDR might act in collaboration<sup>8</sup>, we tested whether other effective shRNAs might increase the effect of *RPS14* knockdown. None of the combinations was more effective than *RPS14* shRNAs alone, suggesting that *RPS14* is the critical gene in the region that explains the haematopoietic differentiation defect associated with 5q<sup>−</sup> syndrome (Supplementary Fig. 4).

To confirm that *RPS14* deficiency truly affects the erythroid differentiation programme (rather than simply modulating the expression of specific FACS markers), we performed genome-wide expression profiling of cells infected with control or *RPS14* shRNAs.

<sup>1</sup>Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. <sup>2</sup>Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>3</sup>Division of Hematology, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>4</sup>Division of Hematology Oncology, University of Massachusetts Medical School, Worcester, Massachusetts 01655, USA. <sup>5</sup>Division of Hematology/Oncology, Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. <sup>6</sup>Department of Biochemistry and Molecular Biology, University of Louisville, Kentucky 40292, USA. <sup>7</sup>Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA.

We used gene set enrichment analysis (GSEA)<sup>9</sup> to assess the effect of *RPS14* knockdown on experimentally derived signatures of erythroid and megakaryocytic differentiation (Fig. 2f). As expected, the gene expression pattern of *RPS14* knockdown showed a significant abrogation of the erythroid differentiation signature ( $P < 0.001$ ; Fig. 2g), in the setting of increased signature of neutrophil and platelet differentiation ( $P < 0.001$  for both; Fig. 2h, i). In addition, *RPS14* shRNAs induced a signature of sensitivity to lenalidomide, the only drug approved by the Food and Drug Administration specifically for MDS patients with 5q deletions<sup>5</sup> (Supplementary Fig. 5). The *RPS14* shRNAs knocked down *RPS14* expression by an average of about 60% in these samples, which is consistent with haploinsufficiency as the cause of these phenotypes. To exclude further the possibility of biallelic inactivation of *RPS14*, we sequenced the *RPS14* gene in 32 MDS patient samples and subjected a subset of these samples to high-density single-nucleotide-polymorphism-based copy number analysis and gene expression profiling. In no case did we detect *RPS14* point mutations, cryptic biallelic deletions or loss of expression (for example, by aberrant methylation; see Supplementary Fig. 6). Taken together, these experiments show that partial loss of function of *RPS14* recapitulates the phenotype of the 5q<sup>-</sup> syndrome.

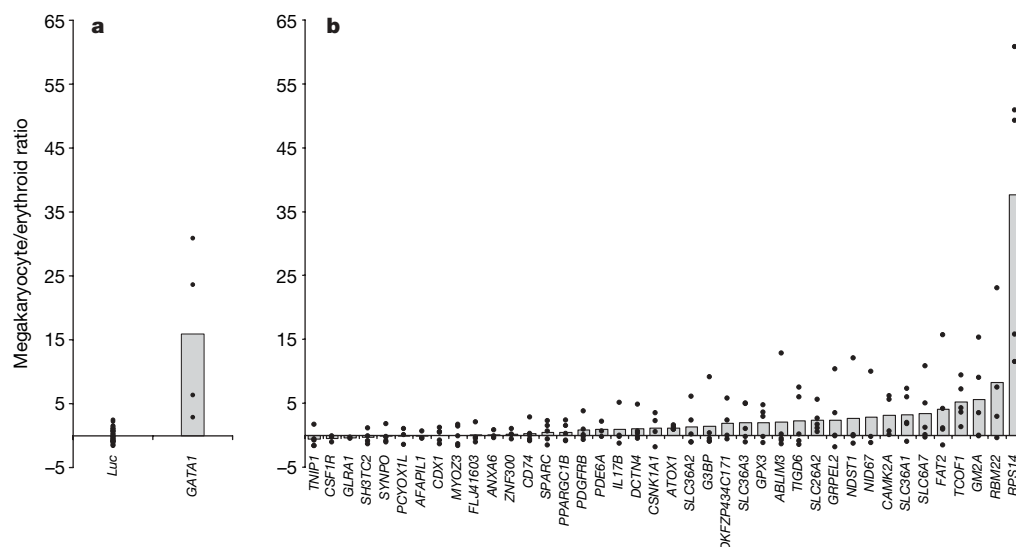
*RPS14* is a component of the 40S ribosomal subunit, but the function of *RPS14* in human cells has not been defined. To determine the effect of partial loss of function of *RPS14* on pre-rRNA processing, we performed northern blotting of rRNA transcripts and sucrose-gradient analysis of intact polysomes. Decreased expression of *RPS14* resulted in an accumulation of the 30S pre-rRNA species with a concomitant decrease in levels of 18S/18SE rRNA levels (Fig. 3a, b), which is consistent with reports for *Saccharomyces cerevisiae* that *RPS14* is required for the processing of 18S pre-rRNA<sup>10</sup>. Specifically, a fourfold to ninefold increase in the 30S/18SE ratio was observed in cells expressing *RPS14* shRNAs. In addition, *RPS14* knockdown abrogated formation of the 40S subunit (Fig. 3c and Supplementary Fig. 7). The increased 30S/18SE ratio in *RPS14*-deficient cells was not simply a consequence of cell death: the ribosomal processing defect occurred before the onset of significant apoptosis, and pharmacologically induced apoptosis failed to generate the characteristic 30S/18S defect (Supplementary Figs 8 and 9).

These results indicate that the block in pre-rRNA processing is a specific consequence of *RPS14* deficiency.

An increase in the 30S/18SE ratio was observed in bone marrow cells from patients with 5q<sup>-</sup> syndrome in comparison with those from normal marrow (Fig. 3d), suggesting that a pre-rRNA processing defect does indeed occur in cells from patients. We note that the samples from patients contain a mixture of normal and 5q<sup>-</sup> disease cells, probably explaining why the 30S/18SE ratio is less perturbed than that seen in the experimental setting. The essential nature of *RPS14* in ribosome biogenesis also probably explains why a complete loss of *RPS14* (for example, through biallelic deletions) is never seen in cells from patients with 5q<sup>-</sup> syndrome. Complete loss of *RPS14* is probably incompatible with cell survival, as it is in yeast<sup>10</sup>.

To establish further that *RPS14* deficiency accounts for the haematopoietic defect characteristic of 5q<sup>-</sup> syndrome, we attempted to rescue the erythroid differentiation defect in patient-derived bone marrow cells by using an *RPS14* expression construct. CD34<sup>+</sup> cells from viably frozen bone marrow mononuclear cells obtained from MDS patients with and without 5q deletions (Supplementary Table 4) were induced to undergo differentiation *in vitro*. FACS analysis showed that in comparison with control, lentiviral expression of *RPS14* increased erythroid differentiation in patients with the 5q<sup>-</sup> syndrome but failed to do so in patients lacking 5q deletions ( $P = 0.004$  for erythroid relative to megakaryocytic differentiation;  $P = 0.0003$  for erythroid relative to myeloid; Fig. 4 and Supplementary Fig. 10). Furthermore, gene expression profiling coupled with GSEA showed that ectopic expression of *RPS14* induced the gene expression signature of erythroid differentiation in 5q<sup>-</sup> syndrome patient samples ( $P < 0.001$ ; Supplementary Fig. 11). These data demonstrate that overexpression of *RPS14* rescues the erythroid differentiation defect seen in patients with 5q<sup>-</sup> syndrome, and establishes *RPS14* as the likely disease-causing gene.

Loss of function of a ribosomal protein might at first seem like an unlikely explanation for a disease with such a distinct haematopoietic phenotype. However, germline heterozygous mutations for two other ribosomal proteins—*RPS19* and *RPS24*—have recently been described in the congenital disorder known as Diamond–Blackfan anaemia<sup>11,12</sup>. The phenotype of Diamond–Blackfan anaemia is strikingly similar to the 5q<sup>-</sup> syndrome: patients have a severe anaemia,



**Figure 1 | Screen of the common deleted region for the 5q<sup>-</sup> syndrome.**

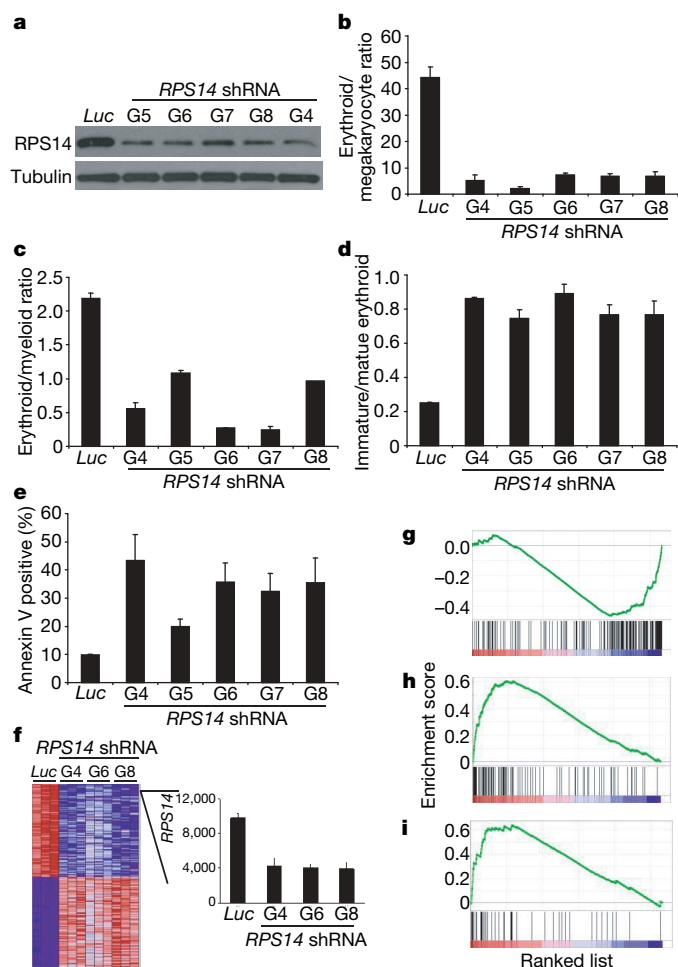
Each gene was targeted by multiple lentivirally expressed shRNAs in CD34<sup>+</sup> cells from umbilical cord blood, and the ratio of megakaryocytic to erythroid differentiation was determined by flow cytometry with antibodies against CD41 and GlyA, respectively. **a**, Controls: an shRNA targeting the luciferase gene (*Luc*), which is not expressed in the primary cells, and multiple shRNAs targeting *GATA1*, encoding an erythroid-specific transcription factor. **b**, All

of the genes in the CDR for the 5q<sup>-</sup> syndrome. The megakaryocytic/erythroid ratio is shown as a z-score using the mean and standard deviation of the control (luciferase) replicates. For the control shRNA targeting the luciferase gene, circles represent 30 individual replicates. For all other genes, circles represent the median of three replicates for each individual shRNA. The mean of all shRNAs targeting a given gene is shown by the height of the grey bar.

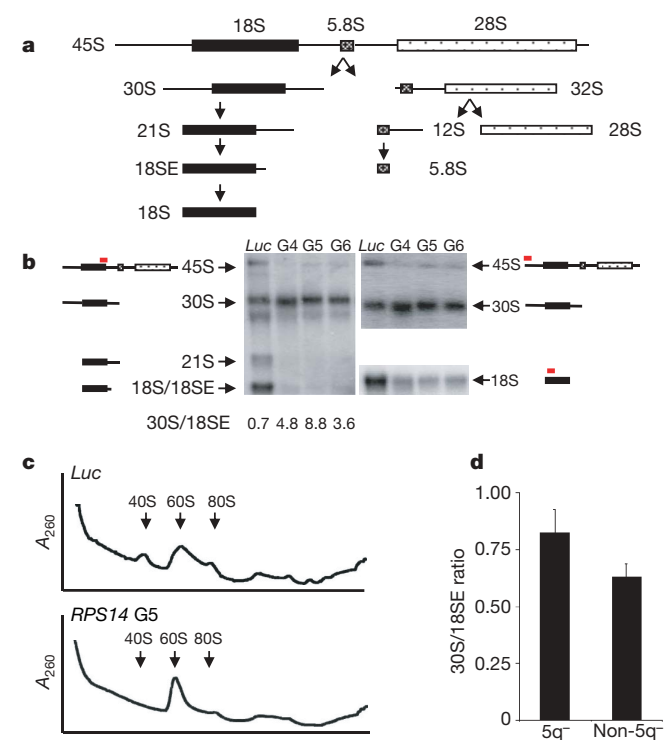
macrocytosis, relative preservation of the platelet and neutrophil counts, erythroid hypoplasia in the bone marrow and an increased risk of leukaemia. Analogous to our results demonstrating RPS14 function in 18S pre-rRNA processing and 40S polysome formation, a similar requirement of RPS19 in ribosomal biogenesis has recently been shown<sup>13</sup>. Beyond Diamond–Blackfan anaemia, the genes implicated in other paediatric bone marrow failure syndromes,

including Shwachman–Diamond syndrome, dyskeratosis congenita and cartilage–hair hypoplasia, are also involved in ribosomal biogenesis<sup>14</sup>. Our findings therefore establish a logical link between the 5q<sup>−</sup> syndrome, caused by the somatic deletion of one allele of *RPS14*, and congenital bone marrow failure syndromes, caused by the heritable mutation of other ribosome-associated proteins.

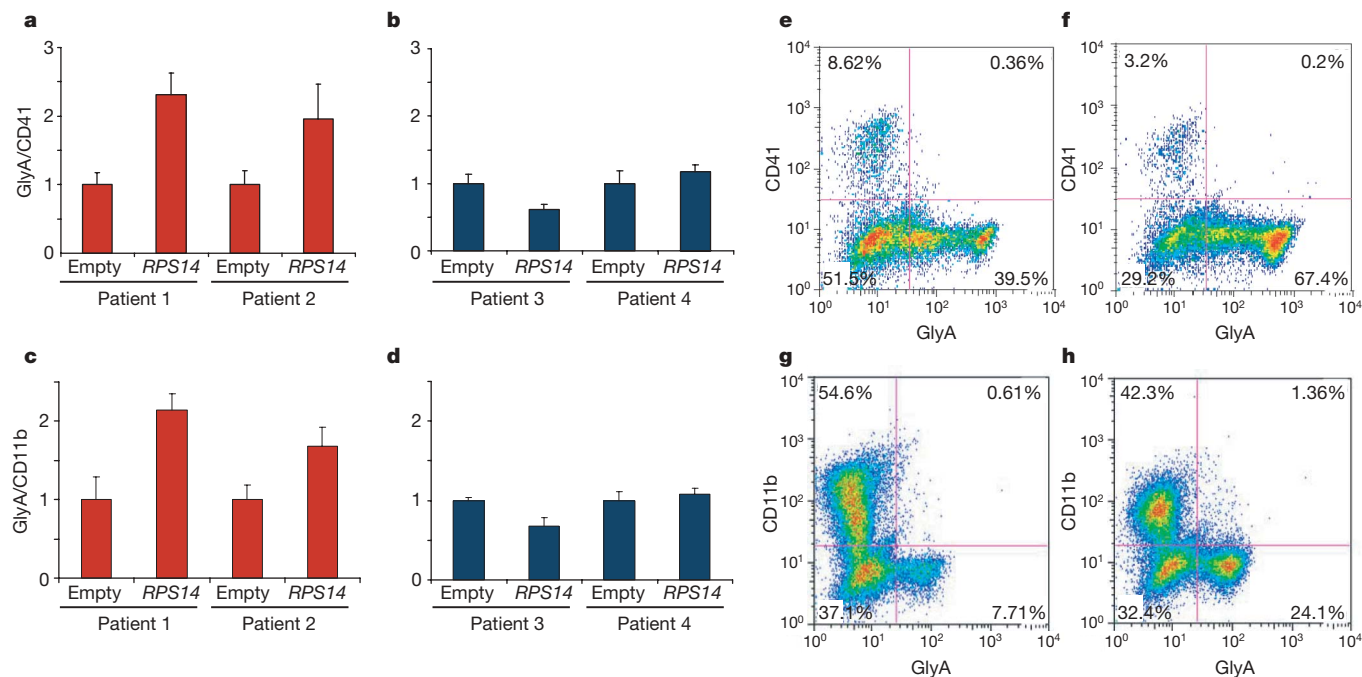
The erythroid specificity of acquired or inherited defects in RPS14, RPS19 or RPS24 expression is noteworthy. Although these ribosomal proteins and the ribosomal subunits they constitute are thought to be ubiquitous, the erythroid lineage is under particularly high biosynthetic demand. Erythroid progenitor cells proliferate extraordinarily rapidly (yielding  $2 \times 10^{11}$  new red cells per day in an adult human)<sup>15</sup>, and contain extremely high concentrations of globin proteins—all resulting in a high demand for ribosomal biogenesis. Furthermore, erythroid cells must balance the production of haem and the translation of globin proteins precisely; otherwise the cells undergo apoptosis<sup>16</sup>. It is therefore possible that partial loss of ribosomal function in other lineages may not result in an obvious phenotype. We also note that in an unbiased screen in zebrafish for genes that cause tumours after the loss of a single allele, 92% of the tumour-prone fish lines had hemizygous mutations in genes encoding ribosomal proteins<sup>17</sup>. These observations suggest that loss of function of *RPS14* in the 5q<sup>−</sup> syndrome may explain not only the erythroid differentiation defect seen in affected patients but also their propensity to progress to acute leukaemia. The mechanism by which ribosomal dysfunction is tumorigenic in fish has yet to be determined.



**Figure 2 | Multiple shRNAs targeting *RPS14* recapitulate the 5q<sup>−</sup> syndrome in vitro.** **a**, Western blots demonstrate that five different shRNAs effectively decrease levels of RPS14. **b**, In comparison with a control shRNA targeting the luciferase gene (*Luc*), each of the five *RPS14* shRNAs blocks erythroid relative to megakaryocytic differentiation in adult bone marrow CD34<sup>+</sup> cells. The ratios of cells from the erythroid and megakaryocytic lineages, indicated on the y axis, were assessed by flow cytometry with antibodies against GlyA and CD41, respectively. **c–e**, In addition, *RPS14* shRNAs decrease erythroid relative to myeloid differentiation, assessed with antibodies against GlyA and CD11b (**c**), block terminal erythroid differentiation, assessed with antibodies against GlyA and CD71 (**d**), and increase apoptosis, assessed by annexin V expression (**e**). In **b–e** the effect of *RPS14* shRNAs, in contrast with the *Luc* shRNA, was statistically significant ( $P < 0.05$  by Student's two-tailed *t*-test, mean and s.e.m. shown ( $n = 3$ )). **f**, Multiple shRNAs targeting *RPS14* also alter the transcriptional programs of lineage-specific differentiation. The top 100 marker genes that are differentially expressed between cells expressing control versus *RPS14* shRNAs, ranked by signal to noise ratio<sup>25</sup>. *RPS14* is at the top of the list of downregulated genes and is expressed at about 40% of the normal level. Error bars indicate s.e.m. **g–i**, *RPS14* shRNAs significantly decrease the expression of an erythroid gene expression signature<sup>26</sup> (**g**) and increase the expression of neutrophil<sup>27</sup> (**h**) and platelet<sup>28</sup> (**i**) signatures, as assessed by GSEA. Genes are ranked by signal/noise ratio according to their differential expression between cells expressing *RPS14* and control shRNAs. Genes in the lineage-specific gene sets are marked with vertical bars, and the enrichment score is shown in green.



**Figure 3 | *RPS14* is required for 18S pre-rRNA processing and 40S ribosomal subunit formation.** **a**, A simplified diagram of pre-rRNA processing. **b**, A defect in the 5' processing of 18S pre-rRNA is evident from northern blots using RNA from TF-1 cells expressing control or *RPS14* shRNAs, with an accumulation of 30S rRNA and a deficiency of 21S and 18SE pre-rRNAs and mature 18S rRNA. The northern blot probes are shown in red. **c**, Polysome profiles from TF-1 cells show that decreased expression of *RPS14* results in a 40S subunit deficiency. **d**, The 30S/18SE pre-rRNA ratio is also increased in RNA from bone marrow mononuclear cells from MDS patients with 5q<sup>−</sup> syndrome ( $n = 4$ ) compared with that from MDS patients without 5q deletions ( $n = 5$ ), as measured by quantification of northern blots ( $P = 0.06$ ). Error bars indicate s.e.m.



**Figure 4 | RPS14 overexpression rescues erythroid differentiation in samples from patients with 5q deletions.** **a–d**, CD34<sup>+</sup> cells from bone marrow aspirates of patients with the 5q<sup>-</sup> syndrome (**a**, **c**; red) and MDS patients without 5q deletions (**b**, **d**; blue) were infected with a lentivirus expressing the RPS14 complementary DNA or an empty vector. In patients with 5q deletions, RPS14 overexpression increased erythroid relative to

megakaryocytic differentiation (**a**, **b**) and erythroid relative to myeloid differentiation (**c**, **d**) shown normalized to the empty-vector control. Means and s.e.m. for three independent experiments are shown. **e–h**, Representative flow cytometry plots for patient 1. In comparison with the empty vector control (**e**, **g**), overexpression of RPS14 (**f**, **h**) results in an increase in GlyA expression and a decrease in CD41 (**e**, **f**) and CD11b (**g**, **h**).

The experiments described here establish RPS14 as a causal gene for the 5q<sup>-</sup> syndrome. However, it is conceivable that other genes (on 5q or elsewhere) collaborate with RPS14 to cause the disease phenotype. We speculate that whereas RPS14 loss of function may be sufficient for the erythroid differentiation defect, additional mutations may be required for RPS14-deficient cells to reach clonal dominance and to progress to malignant transformation to AML. In that regard, the 5q<sup>-</sup> syndrome region on chromosome 5 should be distinguished from a more centromeric locus on 5q that has been associated with therapy-related and aggressive subtypes of MDS as well as AML, and for which two candidate genes have been recently reported<sup>18–20</sup>. In most patients a large portion of 5q is deleted, encompassing both critical regions, so it is possible that the loss of both RPS14 and a second collaborating gene is achieved in a single genetic event.

Acquired deletions are a hallmark of cancer and pre-cancerous states. In general, such deletions flag the existence of a tumour suppressor gene conforming to Knudson's two-hit hypothesis, in which one allele is often deleted and the other allele is inactivated by deletion, mutation or epigenetic modification. However, in multiple tumour types (for example 1p deletions in neuroblastoma, 3p deletions in lung cancer, and 7q deletions in myeloid malignancies) the search for the key tumour suppressor gene has been elusive. A possible explanation for the failure to identify these classic tumour suppressor genes is that oncogenesis is caused by allelic insufficiency<sup>21</sup>. The recent discovery of monoallelic deletions or mutations in PAX5 in acute lymphoblastic leukaemia supports this hypothesis<sup>22</sup>. Our RNA interference-based discovery of the 5q<sup>-</sup> syndrome gene suggests that haploinsufficient disease genes can be identified with this approach. It is possible that the systematic application of RNAi might similarly identify the genes responsible for other diseases caused by allelic insufficiency.

## METHODS SUMMARY

**Culture of haematopoietic progenitor cells.** Primary normal human bone marrow or umbilical cord blood CD34<sup>+</sup> cells were differentiated *in vitro* with a two-phase liquid culture system using combinations of cytokines supporting

erythroid, myeloid and megakaryocytic differentiation<sup>23</sup>. Viable cells from bone marrow aspirates from patients with MDS were collected under a protocol approved by the institutional review board at Massachusetts General Hospital.

**Lentiviral vectors.** Multiple shRNA lentiviruses targeting each gene in the CDR for the 5q<sup>-</sup> syndrome were produced as described previously<sup>24</sup>. The target sequence of each shRNA is listed in Supplementary Table 2.

**Flow cytometry.** Haematopoietic differentiation was assessed by flow cytometry with antibodies specific for terminally differentiated erythroid cells (GlyA), immature erythroid cells (CD71), megakaryocytes (CD41) and myeloid cells (CD11b).

**Microarrays with GSEA.** Linear amplification of RNA was performed with the Ovation kit (Nugen) and labelled cDNA was applied to oligonucleotide microarrays (Affymetrix). GSEA was performed as described previously<sup>9</sup>. Microarray experiments and gene sets are listed in Supplementary Tables 3 and 4, respectively, and the data are available at GEO under accession number GSE9487.

**Ribosomal RNA processing and polysome profiles.** The effect of RPS14 knockdown on pre-rRNA processing was performed by northern blot analysis. Polysome fractionation on a sucrose gradient and spectrophotometric detection were performed as described previously<sup>13</sup>.

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

Received 24 August; accepted 16 November 2007.

- Knudson, A. G. Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl Acad. Sci. USA* **68**, 820–823 (1971).
- Van den Berghe, H. et al. Distinct haematological disorder with deletion of long arm of no. 5 chromosome. *Nature* **251**, 437–438 (1974).
- Heaney, M. L. & Golde, D. W. Myelodysplasia. *N. Engl. J. Med.* **340**, 1649–1660 (1999).
- Giagounidis, A. A., Germing, U. & Aul, C. Biological and prognostic significance of chromosome 5q deletions in myeloid malignancies. *Clin. Cancer Res.* **12**, 5–10 (2006).
- List, A. et al. Lenalidomide in the myelodysplastic syndrome with chromosome 5q deletion. *N. Engl. J. Med.* **355**, 1456–1465 (2006).
- Boulton, J. et al. Narrowing and genomic annotation of the commonly deleted region of the 5q<sup>-</sup> syndrome. *Blood* **99**, 4638–4641 (2002).
- Raza, A. et al. Apoptosis in bone marrow biopsy samples involving stromal and hematopoietic cells in 50 patients with myelodysplastic syndromes. *Blood* **86**, 268–276 (1995).

8. Zender, L. *et al.* Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach. *Cell* **125**, 1253–1267 (2006).
9. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
10. Ferreira-Cerca, S. *et al.* Roles of eukaryotic ribosomal proteins in maturation and transport of pre-18S rRNA and ribosome function. *Mol. Cell* **20**, 263–275 (2005).
11. Drapchinskaia, N. *et al.* The gene encoding ribosomal protein S19 is mutated in Diamond–Blackfan anaemia. *Nature Genet.* **21**, 169–175 (1999).
12. Gazda, H. T. *et al.* Ribosomal protein S24 gene is mutated in Diamond–Blackfan anemia. *Am. J. Hum. Genet.* **79**, 1110–1118 (2006).
13. Flygare, J. *et al.* Human RPS19, the gene mutated in Diamond–Blackfan anemia, encodes a ribosomal protein required for the maturation of 40S ribosomal subunits. *Blood* **109**, 980–986 (2007).
14. Liu, J. M. & Ellis, S. R. Ribosomes and marrow failure: coincidental association or molecular paradigm? *Blood* **107**, 4583–4588 (2006).
15. Quesenberry, P. J. & Colvin, G. A. in *Williams Hematology* 153 (McGraw-Hill, New York, 2005).
16. Quigley, J. G. *et al.* Identification of a human heme exporter that is essential for erythropoiesis. *Cell* **118**, 757–766 (2004).
17. Amsterdam, A. *et al.* Many ribosomal protein genes are cancer genes in zebrafish. *PLoS Biol.* **2**, E139 (2004).
18. Horrigan, S. K. *et al.* Delineation of a minimal interval and identification of 9 candidates for a tumor suppressor gene in malignant myeloid disorders on 5q31. *Blood* **95**, 2372–2377 (2000).
19. Liu, T. X. *et al.* Chromosome 5q deletion and epigenetic suppression of the gene encoding  $\alpha$ -catenin (*CTNNA1*) in myeloid cell transformation. *Nature Med.* **13**, 78–83 (2007).
20. Joslin, J. M. *et al.* Haploinsufficiency of *EGR1*, a candidate gene in the del(5q), leads to the development of myeloid disorders. *Blood* **110**, 719–726 (2007).
21. Fodde, R. & Smits, R. Cancer biology. A matter of dosage. *Science* **298**, 761–763 (2002).
22. Mullighan, C. G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758–764 (2007).
23. Ebert, B. L. *et al.* An RNA interference model of RPS19 deficiency in Diamond–Blackfan anemia recapitulates defective hematopoiesis and rescue by dexamethasone: identification of dexamethasone-responsive genes by microarray. *Blood* **105**, 4620–4626 (2005).
24. Moffat, J. *et al.* A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* **124**, 1283–1298 (2006).
25. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
26. Ebert, B. L. *et al.* An erythroid differentiation signature predicts response to lenalidomide in myelodysplastic syndrome. *PLoS Med.* (in the press).
27. Stegmaier, K. *et al.* Gene expression-based high-throughput screening (GE-HTS) and application to leukemia differentiation. *Nature Genet.* **36**, 257–263 (2004).
28. Gnatenko, D. V. *et al.* Transcript profiling of human platelets using microarray and serial analysis of gene expression. *Blood* **101**, 2285–2293 (2003).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank Broad Institute RNAi and Genetic analysis platforms for advice, single-nucleotide polymorphism analysis and reagents. This work was supported by grants from the National Heart Lung and Blood Institute to T.R.G., B.L.E. and S.R.E. T.R.G. is an investigator of the Howard Hughes Medical Institute.

**Author Contributions** B.L.E., J.P., J.B., C.Y.C., P.T. and S.R.E. performed experiments and analysed data. D.E.R. provided essential reagents. N.G., A.R. and E.A. provided samples from patients. B.L.E. and T.R.G. wrote the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence and requests for materials should be addressed to T.R.G. ([golub@broad.mit.edu](mailto:golub@broad.mit.edu)).

## METHODS

**Culture of haematopoietic progenitor cells.** Cryopreserved human bone marrow CD34<sup>+</sup> cells (Poietics) were obtained from Cambrex. Umbilical cord blood was harvested under a protocol approved by the institutional review board (IRB) at Brigham and Women's Hospital, and CD34<sup>+</sup> cells were purified using CD34<sup>+</sup> MACS microbeads (Miltenyi Biotec). Viable cells from bone marrow aspirates from MDS patients were banked under an IRB-approved protocol at Massachusetts General Hospital. To induce erythroid differentiation, cells were cultured in Serum-Free Expansion Medium (Stem Cell Technologies) supplemented with 100 U ml<sup>-1</sup> penicillin/streptomycin, 2 mM glutamine, 40 µg ml<sup>-1</sup> lipids (Sigma), 100 ng ml<sup>-1</sup> stem cell factor, 10 ng ml<sup>-1</sup> interleukin-3, 10 ng ml<sup>-1</sup> interleukin-6 and 0.5 U ml<sup>-1</sup> erythropoietin. The concentration of erythropoietin was increased to 3 U ml<sup>-1</sup> on day 7. To support both erythroid and megakaryocytic differentiation in a single liquid culture, 50 ng ml<sup>-1</sup> thrombopoietin was added to the culture. To support both erythroid and myeloid differentiation in a single liquid culture, 15 ng ml<sup>-1</sup> granulocyte colony-stimulating factor (Neupogen; Amgen) and 40 ng ml<sup>-1</sup> FLT-3 ligand were added. Cells were harvested for flow cytometry after 10 days of liquid culture.

**Culture of TF-1 cells.** TF-1 cells were maintained in RPMI medium supplemented with 10% fetal bovine serum, 100 U ml<sup>-1</sup> penicillin/streptomycin, 2 mM glutamine and 1 ng ml<sup>-1</sup> granulocyte-macrophage colony-stimulating factor. Doxorubicin and staurosporine were obtained from Calbiochem.

**Lentiviral vectors and infection.** Oligonucleotides encoding shRNAs were cloned into pLKO.1 as described previously<sup>24</sup>. Sequences targeted by each shRNA are listed in Supplementary Table 2. The *RPS14* cDNA was cloned into pLenti6.2/V5-DEST (Invitrogen). Lentiviral backbone vector and packaging plasmids were transfected into 293T cells, and viral supernatant was harvested as described previously<sup>24</sup>. Primary haematopoietic cells were infected with lentivirus one day after being thawed in the presence of 2 µg ml<sup>-1</sup> Polybrene (Sigma) and selected 24 h later with 2 µg ml<sup>-1</sup> puromycin (Sigma) for shRNA lentiviruses, and with 3 µg ml<sup>-1</sup> blasticidin for cDNA-expressing lentiviruses.

**Flow cytometry.** Lineage-specific differentiation was evaluated by flow cytometry. About 5 × 10<sup>5</sup> cells were incubated for 15 min on ice with phycoerythrin, phycoerythrin-Cy5 or fluorescein isothiocyanate-conjugated antibodies against glycophorin-A (CD235a, clone GA-R2; BD Pharmingen), CD71 (clone M-A712; BD Pharmingen), CD11b (clone ICRF44; BD Pharmingen), CD41 (clone HIP8; BD Pharmingen) or annexin V.

**Gene expression profiling.** RNA was purified from mononuclear cells with the use of Trizol (Invitrogen). Linear amplification of 20 ng of total RNA was performed with the Ovation Biotin RNA Amplification and Labelling System (Nugen). Fragmented, labelled cDNA was hybridized to HG\_U133AAofAv2 microarrays (Affymetrix). Raw expression values were normalized by using robust multiarray averaging<sup>29</sup>. Marker genes were ranked with the signal/noise metric<sup>25</sup>; for gene *x* this metric,  $S_x$  is calculated as

$$S_x = (\mu_0 - \mu_1)/(\sigma_0 + \sigma_1)$$

where  $\mu_0$  and  $\sigma_0$  are the mean and standard deviation for gene *x* in class 0, and  $\mu_1$  and  $\sigma_1$  are the respective values for class 1. All microarray experiments are listed in Supplementary Table 3. The complete data set, along with Supplementary Information, is available at <http://www.broad.mit.edu/cancer/pub/5qMDS>.

GSEA was performed as described previously<sup>9</sup>. Erythroid-specific genes were defined by genes that are increased during terminal erythroid differentiation *in vitro*<sup>26</sup>; neutrophil-specific genes were defined by comparing mature neutrophils with primary AML blast cells<sup>27</sup>; a platelet-specific gene set was defined previously<sup>28</sup>; and a lenalidomide signature, developed previously, was defined

by the genes that are expressed at significantly higher levels in bone marrow mononuclear cells from patients who do not respond to lenalidomide, compared with patients who do respond to the drug<sup>26</sup>. Statistical significance was assessed by random permutation of the gene sets<sup>9</sup>. All gene sets are listed in Supplementary Table 4.

**Western blots.** Western blots were performed as described previously, using antibodies against RPS14 (A01; Abnova) at 1:500 dilution and antibodies against  $\alpha$ -tubulin (Ab-2; Neomarkers) at 1:1,000 dilution. Image analysis was performed with ImageJ software (<http://rsb.info.nih.gov/ni-image>).

**Ribosomal RNA analysis.** Total RNA was isolated from TF-1 cells or patient samples by using Trizol (Invitrogen). RNA was fractionated on 1.5% formaldehyde-agarose gels and transferred to Zetaprobe membrane (Bio-Rad). Membranes were washed overnight at 55 °C with 2 × SSC (0.3 M NaCl, 0.03 M sodium citrate, pH 7.0) and 1% SDS and prehybridized for a minimum of 4 h with ULTRAhyb oligonucleotide hybridization buffer (Ambion). The oligonucleotide probes used were as follows: 5' ETS, 5'-ACCGGTCACGACTCGGCA-3' (complementary to sequences 1786–1804 in 5' ETS of the ribosomal RNA transcription unit); 18S, 5'-GCATGGCTTAATCTTTGAGACAAGCATAT-3' (complementary to sequences 3681–2709 in 18S rRNA); and 18S/ITS1, 5'-CCTCGCCCTCCGGGCTCCGTTAATGATC-3' (complementary to sequences 5520–5547 spanning the boundary between 18S rRNA and ITS1). The oligonucleotides, at a concentration of 30 pM, were labelled with [ $\gamma$ -<sup>32</sup>P]ATP by using T4 polynucleotide kinase (New England Biolabs). Membranes were hybridized overnight at 37 °C in ULTRAhyb oligonucleotide hybridization buffer and washed the following morning three times with 6 × SSC at 37 °C. Washed membranes were subjected to phosphorimage analysis (Phosphorimager SF; Molecular Dynamics) for quantification.

**Polysome analysis.** Extracts from TF-1 cells infected with *RPS14* or control shRNAs were prepared as described previously<sup>30</sup>. Extracts were layered on 16-ml 15–55% sucrose gradients and centrifuged in a SW28.1 rotor (Beckman Instruments) for 5 h at 28,000 r.p.m. Gradients were fractionated, and  $A_{254}$  was monitored on an ISCO model 185 gradient fractionator using a UA-6 absorbance detector.

**Statistical analysis.** The significance of experimental results was determined by Student's *t*-test unless otherwise noted. The significance of *RPS14* overexpression relative to control, in samples from patients with 5q deletions compared with patients without 5q deletions, was determined by a two-way analysis of variance.

For the shRNA screen of genes in the 5q CDR, the likelihood that each gene significantly altered differentiation was determined by using a modified Kolmogorov-Smirnov statistic, similarly to the procedure implemented in GSEA<sup>9</sup>. For each gene, the set of shRNAs targeting that gene were combined into a gene set. All scores for the screen were sorted to create a ranked list. The enrichment score of each gene set was calculated by using a modified Kolmogorov-Smirnov statistic. In brief, the enrichment score is computed as a Kolmogorov-Smirnov statistic, namely, the maximum deviation from zero of the difference between the empirical cumulative distribution function (ECDF) of probe scores for a given gene, and the ECDF of the probe scores of all the other genes. Bonferroni *P* values were calculated to correct for multiple hypotheses.

29. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
30. Tang, H. *et al.* Amino acid-induced translation of TOP mRNAs is fully dependent on phosphatidylinositol 3-kinase-mediated signaling, is partially inhibited by rapamycin, and is independent of S6K1 and rpS6 phosphorylation. *Mol. Cell. Biol.* **21**, 8671–8683 (2001).

## LETTERS

# Cyclic dermal BMP signalling regulates stem cell activation during hair regeneration

Maksim V. Plikus<sup>1</sup>, Julie Ann Mayer<sup>1</sup>, Damon de la Cruz<sup>1</sup>, Ruth E. Baker<sup>2</sup>, Philip K. Maini<sup>2,3</sup>, Robert Maxson<sup>4</sup> & Cheng-Ming Chuong<sup>1</sup>

In the age of stem cell engineering it is critical to understand how stem cell activity is regulated during regeneration. Hairs are mini-organs that undergo cyclic regeneration throughout adult life<sup>1</sup>, and are an important model for organ regeneration. Hair stem cells located in the follicle bulge<sup>2</sup> are regulated by the surrounding microenvironment, or niche<sup>3</sup>. The activation of such stem cells is cyclic, involving periodic  $\beta$ -catenin activity<sup>4–7</sup>. In the adult mouse, regeneration occurs in waves in a follicle population, implying coordination among adjacent follicles and the extrafollicular environment. Here we show that unexpected periodic expression of bone morphogenetic protein 2 (*Bmp2*) and *Bmp4* in the dermis regulates this process. This BMP cycle is out of phase with the WNT/ $\beta$ -catenin cycle, thus dividing the conventional telogen into new functional phases: one refractory and the other competent for hair regeneration, characterized by high and low BMP signalling, respectively. Overexpression of noggin, a BMP antagonist, in mouse skin resulted in a markedly shortened refractory phase and faster propagation of the regenerative wave. Transplantation of skin from this mutant onto a wild-type host showed that follicles in donor and host can affect their cycling behaviours mutually, with the outcome depending on the equilibrium of BMP activity in the dermis. Administration of BMP4 protein caused the competent region to become refractory. These results show that BMPs may be the long-sought ‘chalone’ inhibitors of hair growth postulated by classical experiments. Taken together, results presented in this study provide an example of hierarchical regulation of local organ stem cell homeostasis by the inter-organ macro-environment. The expression of *Bmp2* in subcutaneous adipocytes indicates physiological integration between these two thermoregulatory organs. Our findings have practical importance for studies using mouse skin as a model for carcinogenesis, intracutaneous drug delivery and stem cell engineering studies, because they highlight the acute need to differentiate supportive versus inhibitory regions in the host skin.

Mammalian skin contains thousands of hair follicles, each undergoing continuous regenerative cycling. A hair follicle cycles through anagen (growth), catagen (involution) and telogen (resting) phases, and then re-enters the anagen phase. At the base of this cycle is the ability of hair follicle stem cells to briefly exit their quiescent status to generate transient amplifying progeny, but maintain a cluster of stem cells. It is generally believed that a niche microenvironment is important in the control of stem cell homeostasis in various systems<sup>8</sup>. Within a single hair follicle, periodic activation of  $\beta$ -catenin in bulge stem cells is responsible for their cyclic activity<sup>3</sup>. However, how these stem cell activation events are coordinated among neighbouring hairs remains unclear. It is possible that a population of hair follicles

could cycle simultaneously, randomly or in coordinated waves. We recently observed a ‘cyclic alopecia’ phenotype in *Msx2* (homeo box, msh-like 2) null mice, which in essence represents coordinated hair regenerative activity in a population of follicles and is manifest as traversing hair waves<sup>9–11</sup> (Supplementary Fig. 1).

Classical works have documented hair growth waves in rats, mice and other mammals<sup>12,13</sup>. Opinions differ as to whether the hair growth pattern is controlled by local inherent rhythms, systemic factors or both. Because there is a period after anagen during which ‘the systemic stimulus is unable to exert an effect’, the concept of ‘telogen refractivity’ was conceived<sup>14</sup>. A substance, termed ‘chalone’, which can inhibit anagen development, was proposed to explain this phenomenon<sup>15</sup>. However, despite efforts to identify the chalone<sup>16,17</sup>, its molecular nature has remained elusive for the past 50 years.

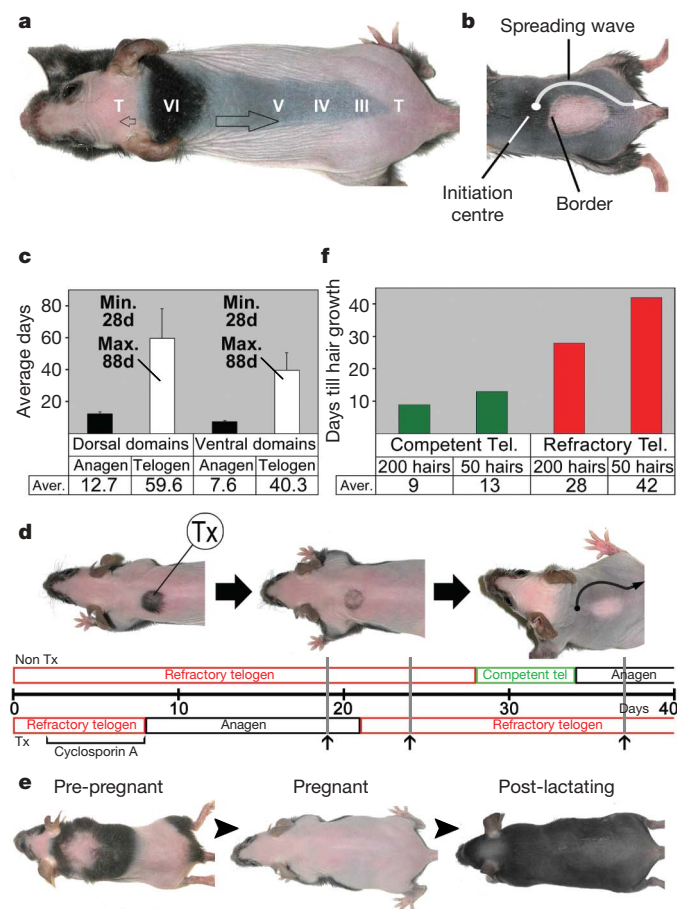
Intrigued by these dynamic, complex hair growth patterns (Supplementary Fig. 1), we set out to find the underlying molecular mechanisms. A hair-cycle domain is a region of skin that contains a population of hair follicles cycling in coordination. The fact that such domains form implies the existence of signals that serve to spread and stop waves of hair growth. This prompted the suggestion that skin regions in telogen can be in either of the two functional phases: competent telogen, which allows the anagen-re-entry wave to propagate, and refractory telogen, which arrests the wave (Fig. 1a, b). We analysed the cycling behaviour of domains in more than 30 living mice (starting from older than 2 months) for up to 1 year (Supplementary Fig. 1), and consistently found that there is a minimal 28-day-long telogen phase; this was defined as early telogen. After this phase, telogen can either end right away (0 days) or persist for any number of days up to about 60 days. This phase (defined as late telogen) contributes to the apparently highly variable telogen length (Fig. 1c).

This suggests that the first 28 days of telogen are essential for the hair cycle and may represent the refractory phase. To test this idea, we used club-hair (a hair filament that has stopped growing but remains attached in the follicle) plucking, which can induce hair regeneration. We gauged responses by the time required for regeneration to start after hairs are plucked (see Methods). When 50 hairs were plucked from skin in the early telogen phase, a longer time was required for hair growth than when a comparable number of hairs were plucked during late telogen (requiring 42 versus 13 days). When 200 hairs were plucked, the time required for hairs to re-grow became shorter but still differed between early and late telogen (28 versus 9 days; Fig. 1f and Supplementary Fig. 2), so anagen re-entry is faster when 200 hairs were plucked versus 50 hairs. Thus, the functional status of a particular skin region can be determined by the hair plucking/regeneration assay. In the follicles we studied, early (up to 28 days)

<sup>1</sup>Department of Pathology, Keck School of Medicine, University of Southern California, Los Angeles, California 90033, USA. <sup>2</sup>Centre for Mathematical Biology, Mathematical Institute, 24–29 St Giles', Oxford, OX1 3LB UK. <sup>3</sup>Oxford Centre for Integrative Systems Biology, Department of Biochemistry, South Parks Road, Oxford OX1 3QU, UK. <sup>4</sup>Department of Biochemistry and Molecular Biology, Keck School of Medicine, University of Southern California, Los Angeles, California 90089, USA.

and late (after 28 days) telogen periods correlate well with refractory and competent telogen phases, respectively.

If the refractory and competent states of hair-cycle domains are transient, then offsetting the timing of hair cycling in a localized



**Figure 1 | Defining refractory and competent telogen.** **a**, Propagation (blank arrows) of hair regenerative waves is seen in *Msx2*-null mice (also see Supplementary Fig. 1). Similar patterns can be seen in normal black mice after hair clipping. Roman characters, anagen stages; T, telogen. **b**, Under physiological conditions, some domains can become refractory to the spreading wave (white arrow). **c**, Normal telogen timing in C57BL/6J mice. The durations of anagen and telogen were measured in 22 hair-cycle domains from dorsal and ventral skin. Error bars represent standard deviation;  $n = 18, 22, 28$  and  $30$ , from left to the right. Min. and Max. represent the range of values, whereas numbers at the bottom represent average (Aver.) number of days. **d**, Experimental induction of refractory telogen with cyclosporine A. The  $x$  coordinate represents the timescale (in days) when experiments began in the early telogen of the non-treated skin region. Cyclosporine A was applied to a localized region (treated, Tx) during early telogen, and induced new anagen about 8 days later. The surrounding non-treated refractory telogen skin (Non Tx) remained in telogen. When the non-treated skin was at day 19 of telogen, treated Tx skin had already proceeded to the late stage of its induced new anagen (left panel, day 19). When non-treated skin was at day 24 of telogen, the cyclosporine-treated region had finished its induced new anagen phase and had entered new telogen (middle panel, day 24). Soon, the non-treated skin progressed into competent telogen. At day 34, the non-Tx region entered its natural anagen. The regenerative wave spread but could not enter the Tx region because it was still in its refractory telogen period (right panel, day 37). Black, anagen; green, competent telogen; red, refractory telogen. **e**, In female mice, multiple hair-cycle domains were reset into one after pregnancy/lactation. Arrowheads, time sequence. **f**, Delayed response to plucking during telogen. Tel., telogen. Hair plucking/regeneration was used to gauge the competent and refractory telogen status ( $n = 16$ ). The minimum time (shown in days) represents the time required for new pigmented hair filaments to be visible. This time is shorter when more hairs were plucked or when the same number of hairs was plucked in the competent period.

region should lead to the formation of new hair-cycle domains. We tested this by local application of cyclosporine A (a powerful anagen-inducing agent that can overcome refractory telogen<sup>18</sup>) to a skin region about 10 mm in diameter that was in telogen day 1. Eight days later, the treated region was in the induced new anagen whereas the surrounding skin continued its progression through refractory telogen. Soon after the treated region completed anagen and re-entered early (new) telogen, the surrounding skin had progressed into late (competent) telogen. When a new hair growth wave approached, it propagated without obstruction over the untreated competent skin, but met resistance in the treated refractory region, thus forming a new hair-cycle domain (Fig. 1d).

Hair-cycle domains are different from regionally specific domains established in development (for example, footpad versus dorsal paw). The exact domain boundaries can shift from cycle to cycle and the domain patterns become more complex as the mouse matures<sup>13</sup> (Supplementary Fig. 1). These complex hair-cycle domains can be affected by systemic factors. For example, during pregnancy and lactation, female mouse hairs that enter telogen are unable to re-enter anagen. Thus, multiple hair-cycle domains are reset into one single domain after pregnancy and lactation<sup>19</sup> (Fig. 1e). Oestrogen and prolactin have been implicated in inhibition of anagen initiation (Supplementary Information).

We wanted to know the molecular mechanisms that constitute this refractoriness. Using *in situ* hybridization and several *lacZ* reporter mice (including *Bmp4-lacZ*, *Nog-lacZ* and the TOPGAL reporter), we searched for cyclic molecular expressions that correlate with refractory and competent telogen. In longitudinal sections of a hair-cycle domain, the hair wave is 'frozen in time' and successive temporal hair-cycle stages are laid out in a spatial order<sup>11</sup>, thus facilitating molecular analyses. We observed canonical WNT signalling and *Msx2*, amongst others, to be expressed in different hair follicle compartments and to fluctuate with hair cycling, as reported (Supplementary Fig. 9). Unexpectedly, we observed the expression dynamics of interfollicular *Bmp2* to be out of phase with that of WNT signalling (Fig. 2a, b, and Supplementary Fig. 5a–e). *Bmp2* expression was absent in early anagen and gradually intensified to reach a peak level in anagen V–VI. *Bmp2* expression remained high in early telogen, but became absent in late telogen (Fig. 2a, b and Supplementary Fig. 5c–e, g). *Bmp4* exhibited similar on and off expression dynamics, as shown by semi-quantitative PCR with reverse transcription, *in situ* hybridization (Fig. 2c, d) and *Bmp4-lacZ* expression (Supplementary Fig. 3). In contrast, *Nog-lacZ* expression showed that on and off dynamics of mesenchymal *Nog* (including dermal papilla and dermal sheath; Supplementary Fig. 4)<sup>17,20</sup> coincides with the hair-cycle rhythm. Because BMP activity can be modulated by multiple factors (different ligands, antagonists and receptors), we measured BMP signalling output by pSMAD (phospho SMAD) 1/5/8 immunostaining; this showed that SMAD 1/5/8 is activated in refractory and is absent in competent telogen hair follicles (Fig. 2e and Supplementary Fig. 6).

We noted that the ability to propagate anagen induction is limited to early anagen follicles. A wave front is halted when it faces a refractory telogen region. By the time this refractory telogen region progresses into competent telogen, the previously propagating anagen follicles have progressed into late anagen and propagation does not resume (Supplementary Fig. 5c, d). Although the surrounding environment is now competent, late anagen follicles are unable to propagate. In this way, the traditional anagen period can be divided into early (anagen I–IV) propagating and late (anagen V, VI) autonomous anagen, with low and high expression of both *Bmp2* and *Bmp4*, respectively, in these phases (Fig. 2a, b and Supplementary Figs 3 and 5). We summarize the rhythms of marker gene expression in Fig. 2g and Supplementary Fig. 10.

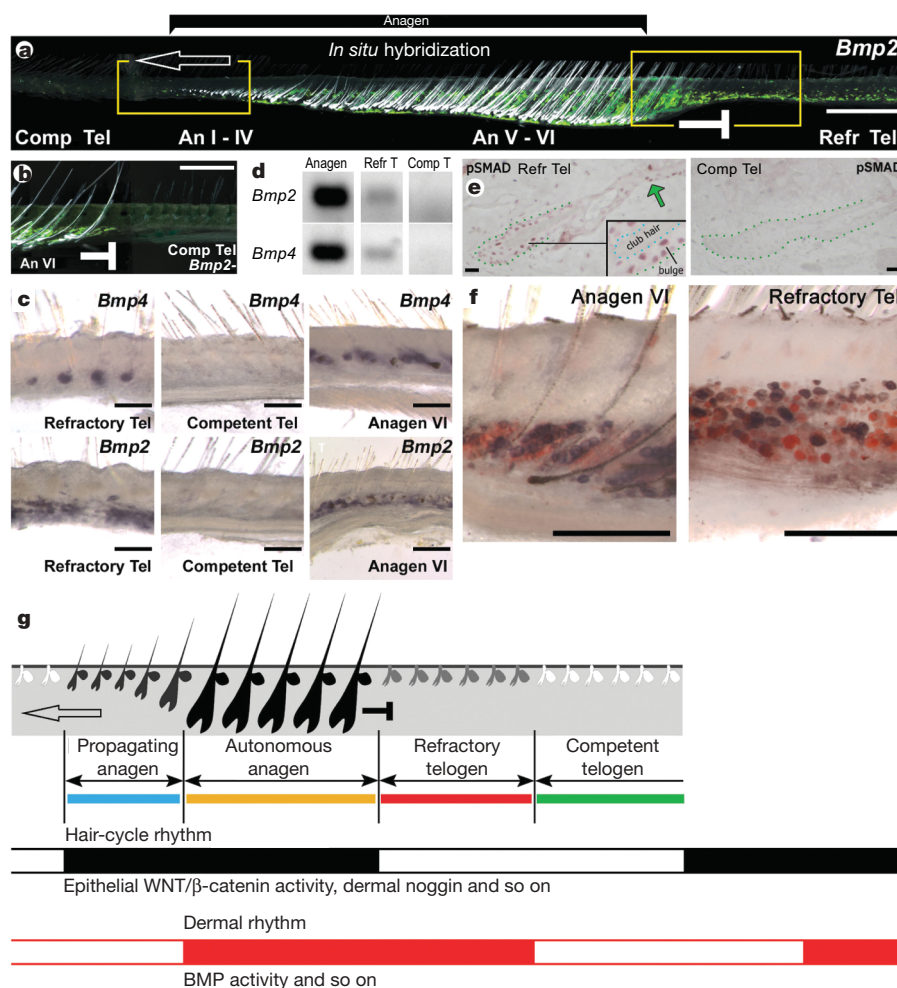
Where are the BMP-producing cells? Most of the periodically expressed *Bmp2* transcripts are produced by subcutaneous adipocytes, as judged by double-staining with Sudan Red (Fig. 2f).

Periodic expression of *Bmp4* is seen in the intrafollicular epithelium, secondary hair germ cells, dermal papilla and adjacent extra-follicular dermal fibroblasts (Supplementary Fig. 4). Collectively, we define the extrafollicular sources of periodic *Bmp2* and *Bmp4* expression as the dermal macroenvironment. The macroenvironmental BMPs may have a large additive effect on the strength of intrafollicular (microenvironmental) BMP6 and BMP4 signalling<sup>21,22</sup> in regulating the quiescence of pSMAD-positive bulge stem cells, although these mechanisms remain to be investigated. Because the eventual anagen initiation requires activation of WNT/ $\beta$ -catenin<sup>3,4</sup>, there is competitive equilibrium between BMP and WNT signalling<sup>21</sup>. Stem cells have to integrate the multiple signalling inputs from both the microenvironment and the macroenvironment to make the decision.

The first telogen (around postnatal day 19) is very short and new anagen initiates quickly without detectable refractory telogen. Dermis acquires telogen refractivity with maturation and second telogen (postnatal day 45–70; Supplementary Fig. 6b) does have refractory telogen. These findings lead us to hypothesize that: first, in the ‘BMP on’ phase, the macroenvironment prevents microenvironment-based activation of bulge stem cells (by means of

WNT signalling), resulting in refractory telogen; and, second, in the ‘BMP off’ phase, the macroenvironmental block is removed and the threshold for microenvironment-based activation of stem cells is low. This results in competent telogen; hair follicles are free to enter new anagen either by stochastic self-activation or by facilitation by adjacent early anagen follicles. We tested this hypothesis by transgenic perturbation of BMP signalling, skin transplantation and administration of exogenous BMP4.

If BMPs have a causative role in conferring refractory status, we should be able to reduce the period of refractory telogen by down-regulating BMP signalling. We did this by overexpressing *Nog* under the keratin 14 promoter in *Krt14–Nog* mice<sup>23</sup> (named K14–Noggin in ref. 23). The minimal telogen length was reduced to 6 days, and the maximal length was reduced to 11 days (Fig. 3b). As a result, these mice displayed continuous propagation of hair regenerative waves and have highly simplified hair-cycle-domain patterns (Fig. 3a). We further tested the response of *Krt14–Nog* hair follicles to hair plucking. The differences in response we observed in wild-type mice in early versus late telogen were eliminated in *Krt14–Nog* mice. In all cases, plucked *Krt14–Nog* hair follicles required only approximately 6 days to re-enter anagen (Fig. 3c). Recently, the importance of BMP



**Figure 2 | Periodic BMP signalling in the dermis and subcutaneous adipose tissue.** **a**, Different temporal stages are spatially laid across the skin strip. The dark-field illumination shows hair follicles (white) and *Bmp2* *in situ* hybridization (green). Note that the beginning and end of the hair cycle and the beginning and end of *Bmp2* *in situ* are out of phase. An, anagen; Comp, competent; Refr, refractory; Tel, telogen. Open arrow, the direction of the spreading waves; stop sign, boundary between anagen and refractory telogen. **b**, When the refractory telogen region becomes competent, anagen VI follicles still do not propagate. **c**, **d**, *Bmp2* and *Bmp4* expressions are

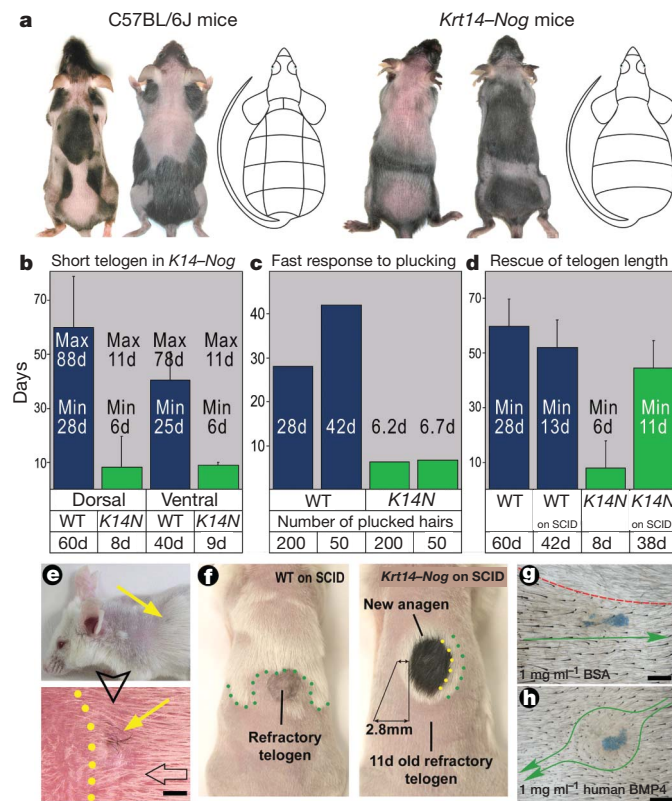
detected by *in situ* and semi-quantitative PCR with reverse transcription. Both methods show that *Bmp2* and *Bmp4* are present in late anagen and refractory telogen, but absent in competent telogen. **e**, pSMAD immunostaining is present in follicular epithelium, including in the bulge area (inset) and adjoining infundibulum (green arrow). **f**, *Bmp2* expression (blue) colocalized within some Sudan-Red-positive adipocytes (red). **g**, Schematic summary of the hair-cycle rhythm (black) and the newly identified dermal rhythm (red). Together, they define four new functional stages. Catagen is omitted for simplification. Scale bars: **a**, 1 mm; **b**, 500  $\mu$ m; **c**, **e**, **f**, 200  $\mu$ m.

activity in suppressing stem cell activity has also been shown by tissue-specific deletion of BMP receptors<sup>21,24</sup>.

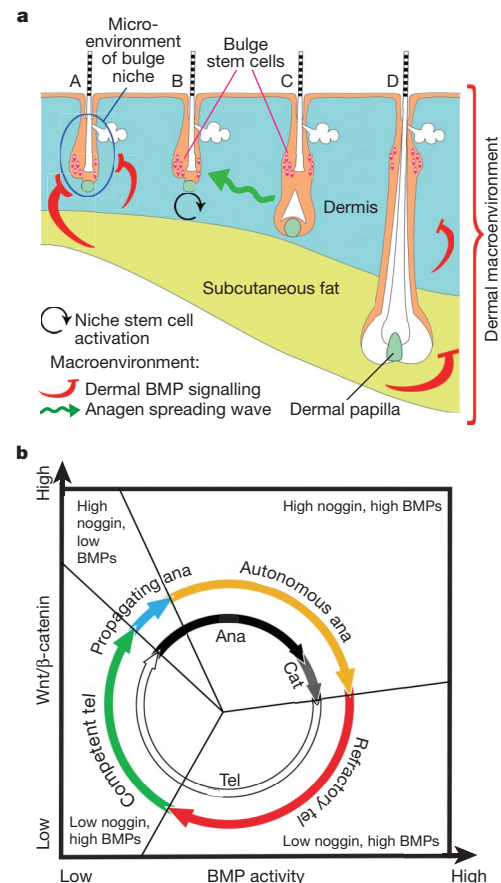
The currently held concept of the stem cell microenvironment implies only autonomous regulation: thus, the activation of stem cells depends only on signalling inputs from components intrinsic to the organ (here, the hair follicle itself<sup>3</sup>). To test directly whether the activation of stem cells is also subjected to non-autonomous regulation, we transplanted skin grafts from pigmented *Krt14-Nog* mice onto albino severe combined immunodeficient (SCID) mice. If the control of stem cell activation is intrinsic to the follicles, hair cycling behaviour should remain the same for both donor and host. Instead, we observed donor–host interactions, reflecting a non-autonomous relationship, with the outcome dependent on the size of the transplanted skin graft. When a small graft of *Krt14-Nog* skin (~1 mm) was transplanted, the donor skin remained in telogen for longer and

could respond to an anagen-activating wave originating from the host (Fig. 3e and Supplementary Fig. 7). Thus, we achieved partial functional rescue of *Krt14-Nog* phenotypes. In contrast, when a large skin graft (>10 mm) was transplanted, the graft exhibited a greater degree of autonomous control within itself. Host telogen hair follicles surrounding the graft re-entered anagen (visible as a rim of white hairs) when pigmented donor hairs entered anagen (Fig. 3f) after only 11 days in telogen (versus 28 days), thus providing evidence of a donor effect on the host.

Classical experiments using skin graft transplantation to ask whether hair growth patterns are controlled intrinsically or systemically have produced variable results<sup>14</sup>. We repeated autologous skin transplantation experiments and observed that hair growth patterns are initially intrinsic to the donor but gradually become entrained to the host rhythm after several hair cycles (not shown). Consequently, the discrepancy amongst classical experiments may be due to the size of the graft and the time they chose for readout. At the molecular level, our results demonstrate involvement of the BMP pathway in the non-autonomous interactions among follicle populations. It remains to be investigated whether the process depends on the direct diffusion of BMPs or their antagonists, or whether it is indirectly mediated by other mechanisms<sup>25</sup>.



**Figure 3 | Altered hair regenerative wave dynamics in *Krt14-Nog* mice, and non-autonomous interactions with normal cycling host skin after transplantation.** **a**, Control (left) and *Krt14-Nog* (right) mice. Hair-cycle domains in two different stages are shown, together with schematic domain boundaries generated by similar analysis to that used in Supplementary Fig. 1. **b**, Measurements show that both refractory and competent telogen are shortened in *Krt14-Nog* mice (*K14N*, green bars) compared to wild type (WT, blue bars). In **b** and **d**, Min and Max represent range of values, whereas numbers at the bottom represent average number of days. In **c**, however, numbers at the bottom represent numbers of plucked hairs. Error bars, standard deviation;  $n = 71$  for *Nog* mice and  $n = 22$ –30 for the control. **c**, Plucking/regenerative response in *Krt14-Nog* (green bars) is about 5 times faster. **d**, **e**, When a small *Krt14-Nog* skin graft was transplanted into SCID skin, hair growth (**e**) and duration of refractory telogen (**d**) were partially rescued (error bars, standard deviation;  $n > 15$ ). The yellow dotted line represents the anagen wave front. Yellow arrows point at the transplanted *Krt14-Nog* hair follicles. The blank arrow points at the spreading direction of the anagen wave. The blank arrowhead points at the enlarged view of the top panel. **f**, When a large *Krt14-Nog* skin graft (>10 mm) was transplanted, it caused reduction of refractory telogen by inducing a rim of white hair in the host. **g**, **h**, Human-BMP4-soaked beads caused hair propagation wave (green arrow) to go around them, creating a new telogen domain. Albumin does not have this effect. Red dashed line, domain border. Scale bars: **e**, **g**, **h**, 1 mm.



**Figure 4 | Functional phases of the hair cycle.** **a**, Illustration of the bulge niche microenvironment and interfollicular dermal macroenvironment, including dermis, subcutaneous fat and adjacent follicles. Anagen-stimulating (black and green) or -inhibiting (red) activities are depicted with coloured arrows. Follicles are in different stages: A, refractory telogen; B, competent telogen; C, propagating anagen; and D, autonomous anagen follicles. Blue circle in A, intrafollicular microenvironment; colour-coded similar to panel **b**. **b**, New functional phases (coloured outer circle) mapped against classical hair-cycle stages (black and white inner circle). On the basis of the growth-inducing ability of the follicles, anagen is divided into propagating (inducing blue) and autonomous (non-inducing, yellow) phases. On the basis of the ability to respond to regenerative signals, telogen is divided into refractory telogen (red) and competent (green) phases.

Finally, we tested whether a direct local delivery of BMP protein can convert competent telogen status to refractory in normal mice. Human-BMP4-soaked beads were implanted into competent telogen skin ahead of an anagen-spreading wave (see Methods<sup>17</sup>). Twelve days later, human BMP4, but not control BSA, prevented the propagation of the wave around the beads (Fig. 3g, h and Supplementary Fig. 8). Thus, the level of BMP activity can indeed explain the functional status (refractory versus competent) of a skin region.

Results here add new dimensions to our understanding of skin biology. First, these findings demonstrate that, in addition to short distance microenvironmental control<sup>17,22</sup>, the activation of stem cells within large groups of hair follicles is subject to long distance macro-environmental control from the surrounding dermis (Fig. 4). This concept is readily applicable to other organs. For example, whereas *Bmp4* is constantly expressed in the mesenchyme of intestinal microvilli, bursts of *Nog* expression in the villi stem cell niche may act transiently to lower BMP signalling, thus allowing stem cells to proliferate for epithelial renewal<sup>26</sup>. Second, extrafollicular periodically expressed *Bmp2* and *Bmp4* seem to fulfil the criteria of the elegant but elusive chalone proposed to explain patterned hair growth<sup>14,15,17</sup>, thus solving a 50-year-old puzzle. Third, the dynamic expression of *Bmp2* in dermal adipocytes suggests a link between two skin organ systems. Because subcutaneous fat, like hairs, has a thermo-regulatory function and leptin is present in the dermal papilla of hair follicles<sup>27</sup>, periodically expressed *Bmp2* may coordinate the function of these two organs in response to the external environment and may have implications for the evolution of integuments<sup>28</sup>. Fourth, the asynchronous cyclic expression of BMPs and  $\beta$ -catenin in the dermis and hair follicle provide a platform for mutual modulations of these 'clocks' in the skin. They also imply that stem cell regeneration is subject to the control of biological rhythms.

Finally, mouse skin has been used extensively as a model in studies of carcinogenesis, intra-cutaneous drug delivery and stem cell biology<sup>29,30</sup>. Such studies are usually designed on the assumption that the skin is a stable and largely uniform medium. Our findings show clearly that this assumption is rarely, if ever, justified.

## METHODS SUMMARY

**Animals.** C75BL/6J, Crl:CD1(ICR), C3H/HeJ and SCID mice were used in this study. *Msx2* null (C.Cg-*Msx2*<sup>tm1Rim</sup>/Mmcd), *Krt14-Nog* (B6.CBA-Tg(*Krt14-Nog*)), *Bmp4-lacZ* (129S-*Bmp4*<sup>lacZneo</sup>), *Nog-lacZ* (129S-*Nog*<sup>tm1Amc</sup>/J) and TOPGAL (STOCK Tg(*Fos-lacZ*)34Efu/J) transgenic mice were also used.

**Hair-cycle observation.** Progression of hair growth patterns was monitored in mice for various intervals of time, up to 1 year. Hair clipping was selected over plucking or shaving to avoid wounding that can potentially interfere with normal hair growth<sup>13,15</sup>.

**Animal procedures.** All procedures were performed on anaesthetized animals with protocols approved by USC vivaria. For skin transplantation, surgical procedures were performed when both donor and recipient skins were in early telogen. This was done to ensure that wounded skin is healed by the beginning of the next anagen phase and that the affect of wound healing on the hair cycle is minimal. SCID mice were used as recipients.

**Histology and detection of molecular expressions.** Tissues were collected, fixed and processed for histology as described<sup>13,23</sup>.

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

Received 9 July; accepted 7 November 2007.

1. Stenn, K. S. & Paus, R. Controls of hair follicle cycling. *Physiol. Rev.* **81**, 449–494 (2001).
2. Morris, R. J. *et al.* Capturing and profiling adult hair follicle stem cells. *Nature Biotechnol.* **22**, 411–417 (2004).
3. Fuchs, E., Tumber, T. & Guasch, G. Socializing with the neighbors: stem cells and their niche. *Cell* **116**, 769–778 (2004).
4. Huelsken, J., Vogel, R., Erdmann, B., Cotsarelis, G. & Birchmeier, W.  $\beta$ -Catenin controls hair follicle morphogenesis and stem cell differentiation in the skin. *Cell* **105**, 533–545 (2001).

5. Reddy, S. *et al.* Characterization of Wnt gene expression in developing and postnatal hair follicles and identification of Wnt5a as a target of Sonic hedgehog in hair follicle morphogenesis. *Mech. Dev.* **107**, 69–82 (2001).
6. Lo Celso, C., Prowse, D. M. & Watt, F. M. Transient activation of  $\beta$ -catenin signalling in adult mouse epidermis is sufficient to induce new hair follicles but continuous activation is required to maintain hair follicle tumours. *Development* **131**, 1787–1799 (2004).
7. Lowry, W. E. *et al.* Defining the impact of  $\beta$ -catenin/Tcf transactivation on epithelial stem cells. *Genes Dev.* **19**, 1596–1611 (2005).
8. Moore, K. A. & Lemischka, I. R. Stem cells and their niches. *Science* **311**, 1880–1885 (2006).
9. Ma, L. *et al.* 'Cyclic alopecia' in *Msx2* mutants: defects in hair cycling and hair shaft differentiation. *Development* **130**, 379–389 (2003).
10. Militzer, K. Hair growth pattern in nude mice. *Cell. Tiss. Org.* **168**, 285–294 (2001).
11. Suzuki, N., Hirata, M. & Kondo, S. Traveling stripes on the skin of a mutant mouse. *Proc. Natl Acad. Sci. USA* **100**, 9680–9685 (2003).
12. Durward, A. & Rudall, K. M. Studies on hair growth in the rat. *J. Anat.* **83**, 325–335 (1949).
13. Plikus, M. V. & Chuong, C. M. Complex hair cycle domain patterns and regenerative hair waves in living rodents. *J. Invest. Dermatol.* (in the press) (2007).
14. Ebling, F. J. & Johnson, E. Systemic influence on activity of hair follicles in skin homografts. *J. Embryol. Exp. Morphol.* **9**, 285–293 (1961).
15. Chase, H. Growth of the hair. *Physiol. Rev.* **34**, 113–126 (1954).
16. Paus, R., Stenn, K. S. & Link, R. E. Telogen skin contains an inhibitor of hair growth. *Br. J. Dermatol.* **122**, 777–784 (1990).
17. Botchkarev, V. A. *et al.* Noggin is required for induction of the hair follicle growth phase in postnatal skin. *FASEB J.* **15**, 2205–2214 (2001).
18. Maurer, M., Handjiski, B. & Paus, R. Hair growth modulation by topical immunophilin ligands: induction of anagen, inhibition of massive catagen development, and relative protection from chemotherapy-induced alopecia. *Am. J. Pathol.* **150**, 1433–1441 (1997).
19. Johnson, E. Quantitative studies of hair growth in the albino rat. II. The effect of sex hormones. *J. Endocrinol.* **16**, 351–359 (1958).
20. Botchkarev, V. A. *et al.* Noggin is a mesenchymally derived stimulator of hair-follicle induction. *Nature Cell Biol.* **1**, 158–164 (1999).
21. Kobiak, K., Stokes, N., de la Cruz, J., Polak, L. & Fuchs, E. Loss of a quiescent niche but not follicle stem cells in the absence of bone morphogenetic protein signaling. *Proc. Natl Acad. Sci. USA* **104**, 10063–10068 (2007).
22. Blanpain, C., Lowry, W. E., Geoghegan, A., Polak, L. & Fuchs, E. Self-renewal, multipotency, and the existence of two cell populations within an epithelial stem cell niche. *Cell* **118**, 635–648 (2004).
23. Plikus, M. *et al.* Morpho-regulation of ectodermal organs: integument pathology and phenotypic variations in K14-Noggin engineered mice through modulation of bone morphogenetic protein pathway. *Am. J. Pathol.* **164**, 1099–1114 (2004).
24. Zhang, J. *et al.* Bone morphogenetic protein signaling inhibits hair follicle anagen induction by restricting epithelial stem/progenitor cell activation and expansion. *Stem Cells* **24**, 2826–2839 (2006).
25. Oro, A. E. & Higgins, K. Hair cycle regulation of Hedgehog signal reception. *Dev. Biol.* **255**, 238–248 (2003).
26. He, X. C. *et al.* BMP signaling inhibits intestinal stem cell self-renewal through suppression of Wnt- $\beta$ -catenin signaling. *Nature Genet.* **36**, 1117–1121 (2004).
27. Iguchi, M., Aiba, S., Yoshino, Y. & Tagami, H. Human follicular papilla cells carry out nonadipose tissue production of leptin. *J. Invest. Dermatol.* **117**, 1349–1356 (2001).
28. Wu, P. *et al.* Evo-Devo of amniote integuments and appendages. *Int. J. Dev. Biol.* **48**, 249–270 (2004).
29. Sausville, E. A. & Burger, A. M. Contributions of human tumor xenografts to anticancer drug development. *Cancer Res.* **66**, 3351–3354 (2006).
30. Zheng, Y. *et al.* Organogenesis from dissociated cells: generation of mature cycling hair follicles from skin-derived cells. *J. Invest. Dermatol.* **124**, 867–876 (2005).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank V. Botchkarev, G. Cotsarelis, B. Morgan, R. Paus, J. Sundberg and R. Widelitz for discussions. We are grateful to B. Hogan, R. Harland and S. Bellusci for providing transgenic mice. This work is supported by Grants from NIAMS and NIA from the NIH, USA, to C.-M.C. M.V.P. is a postdoctoral scholar of the California Institute of Regenerative Medicine. R.E.B. is supported by a Research Councils UK Fellowship and a Microsoft European Postdoctoral Research Fellowship.

**Author Contributions** M.V.P. and C.-M.C. designed the experiment and analysed results together. M.V.P. did major bench work and observations. J.A.M. and D.d.I.C. helped with some bench work. R.E.B. and P.K.M. helped to develop the model. R.M. helped by providing mice and discussing the results.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence and requests for materials should be addressed to C.-M.C. (cmchuong@usc.edu).

## METHODS

**Choosing early versus late telogen skin.** To choose early versus late telogen skin in living mice, we used the following protocol.

First, an area on the adult mouse skin where hairs appeared to be growing was chosen. The use of pigmented mice made it easier to distinguish these phases. Hairs were clipped (not plucked) near the skin surface. Anagen-phase skin contains pigment in the proximal hair follicles. This determination can be aided by observing the skin under a dissection microscope, especially when the skin is wet with saline solution to make it appear transparent. These mice were monitored daily, and the day on which skin pigmentation ceased was recorded. This coincides with the anagen/catagen junction. We then waited for an additional 5 days to ensure that skins are in early telogen, giving us early telogen skin to work with. Alternatively, we waited for at least 40 days (well over 4 weeks) after the anagen/catagen junction for late telogen skin to develop, giving us late telogen skin to work with.

**Scoring the plucking experiments.** Hairs were plucked from the early or late telogen region. After plucking, each plucked spot was monitored daily under a dissection microscope. We were able to detect new anagen skin on living mice without having to biopsy or kill the mice for histological specimens. We then looked for changes in pigmentation since the start of melanogenesis in anagen III. Pigmented hairs can be spotted under a dissection microscope before the new hair fibres reach the skin surface. Thus, we were able to record non-invasively the appearance of anagen III hair follicles (when we spotted black hairs under the skin surface). Approximately, this corresponds to the second day of new anagen. It takes another day for the new hair fibre to reach the skin surface. Thus, we were also able to record non-invasively day-3 anagen follicles when the new hair filaments reach above the skin surface.

Because the changes in skin pigmentation are not easily visible, we used the appearance of new hair filaments above the skin surface as the criteria for scoring hair-plucking experiments. Therefore, the results shown in Fig. 1f indicate that it takes approximately 9 days to observe the appearance of day-3 anagen follicles. The extra time includes the period required for the follicle to heal and to get ready to enter anagen.

**Protein administration experiment.** Intracutaneous administration of exogenous protein was performed as follows. Affinity chromatography Affi-gel blue gel beads were obtained from Biorad. Beads were washed in  $1\times$  PBS, followed by drying. The beads were then re-suspended in  $5\text{ }\mu\text{l}$  protein solution, either control (BSA  $1\text{ mg ml}^{-1}$ ) or experimental (human BMP4  $1\text{ mg ml}^{-1}$ ), at  $4^\circ\text{C}$  for 30 min. Recombinant human BMP4 protein was obtained from R&D Systems. Reconstitution of the protein was performed in  $4\text{ mM HCl}$  in  $0.2\%$  BSA as per the manufacturer's guidelines. Approximately 100 beads were introduced to the competent telogen skin of adult mice by means of a single puncture wound to the skin made by a 30 g syringe (insulin syringe). To replenish proteins, subsequent doses of  $1.5\text{ }\mu\text{l}$  protein solution were microinjected to the site of the bead implantation every 24 h by means of a glass micro-needle until the tissue was harvested. After we noted the anagen-spreading wave pass beyond the bead implantation sites (1 week in the case of Fig. 3g, h), we collected the skin and inverted it for photography. This allows the study of the anagen-wave-spreading dynamics around the control and human BMP4 beads.

# Identification of cells initiating human melanomas

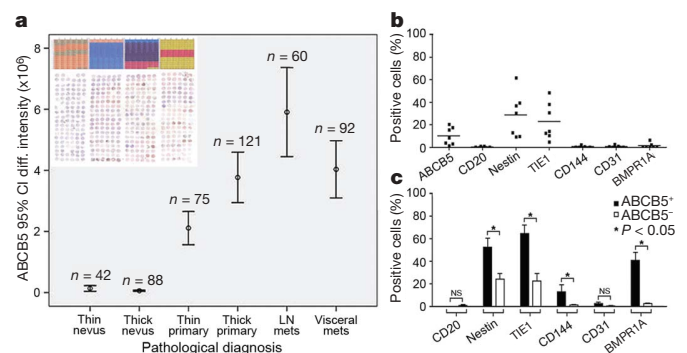
Tobias Schatton<sup>1</sup>, George F. Murphy<sup>2</sup>, Natasha Y. Frank<sup>1,3</sup>, Kazuhiro Yamaura<sup>1</sup>, Ana Maria Waaga-Gasser<sup>4</sup>, Martin Gasser<sup>4</sup>, Qian Zhan<sup>2</sup>, Stefan Jordan<sup>1</sup>, Lyn M. Duncan<sup>5</sup>, Carsten Weishaupt<sup>6</sup>, Robert C. Fuhlbrigge<sup>6</sup>, Thomas S. Kupper<sup>6</sup>, Mohamed H. Sayegh<sup>1</sup> & Markus H. Frank<sup>1</sup>

Tumour-initiating cells capable of self-renewal and differentiation, which are responsible for tumour growth, have been identified in human haematological malignancies<sup>1,2</sup> and solid cancers<sup>3–6</sup>. If such minority populations are associated with tumour progression in human patients, specific targeting of tumour-initiating cells could be a strategy to eradicate cancers currently resistant to systemic therapy. Here we identify a subpopulation enriched for human malignant-melanoma-initiating cells (MMIC) defined by expression of the chemoresistance mediator ABCB5 (refs 7, 8) and show that specific targeting of this tumorigenic minority population inhibits tumour growth. ABCB5<sup>+</sup> tumour cells detected in human melanoma patients show a primitive molecular phenotype and correlate with clinical melanoma progression. In serial human-to-mouse xenotransplantation experiments, ABCB5<sup>+</sup> melanoma cells possess greater tumorigenic capacity than ABCB5<sup>−</sup> bulk populations and re-establish clinical tumour heterogeneity. *In vivo* genetic lineage tracking demonstrates a specific capacity of ABCB5<sup>+</sup> subpopulations for self-renewal and differentiation, because ABCB5<sup>+</sup> cancer cells generate both ABCB5<sup>+</sup> and ABCB5<sup>−</sup> progeny, whereas ABCB5<sup>−</sup> tumour populations give rise, at lower rates, exclusively to ABCB5<sup>−</sup> cells. In an initial proof-of-principle analysis, designed to test the hypothesis that MMIC are also required for growth of established tumours, systemic administration of a monoclonal antibody directed at ABCB5, shown to be capable of inducing antibody-dependent cell-mediated cytotoxicity in ABCB5<sup>+</sup> MMIC, exerted tumour-inhibitory effects. Identification of tumour-initiating cells with enhanced abundance in more advanced disease but susceptibility to specific targeting through a defining chemoresistance determinant has important implications for cancer therapy.

Human malignant melanoma is a highly aggressive and drug-resistant cancer<sup>9</sup> that shows tumour heterogeneity<sup>10,11</sup> and contains cancer cell subsets with enhanced tumorigenicity<sup>12,13</sup>. We predicted that the melanoma chemoresistance mediator ABCB5 (refs 7, 8) could represent a molecular marker defining tumorigenic MMIC, because its expression also characterizes progenitor cell subsets in physiological skin<sup>14</sup>.

We first examined the relationship of ABCB5 to clinical malignant melanoma progression because of its close association with CD166 (ref. 7), a marker of more advanced disease<sup>15</sup>. This was assessed by ABCB5 immunohistochemical staining of an established melanoma progression tissue microarray<sup>16</sup> representing four major diagnostic tumour types: benign melanocytic nevi, primary cutaneous melanoma, metastases to lymph nodes and metastases to viscera. We found that primary or metastatic melanomas expressed significantly more ABCB5 than benign melanocytic nevi, thick primary melanomas more than thin primary melanomas, and melanomas metastatic to

lymph nodes more than primary lesions (Fig. 1a), identifying ABCB5 as a molecular marker of neoplastic progression. Apparent heterogeneity in ABCB5 expression was noted in metastases, with greater staining in the lymph node than in visceral metastases. When assayed in single-cell suspensions derived from clinical melanomas (Supplementary Table 1), ABCB5 was also found to be consistently expressed in 7/7 specimens, with ABCB5<sup>+</sup> tumour cell frequency ranging from 1.6 to 20.4% ( $10.1 \pm 2.9\%$ , mean  $\pm$  s.e.m.) (Fig. 1b, and Supplementary Table 1). Further characterization with respect to antigens indicative of a more primitive melanoma phenotype revealed expression of CD20 (also known as MS4A1)<sup>12</sup> in 4/7 specimens (cell frequency  $0.4 \pm 0.2\%$ , mean  $\pm$  s.e.m.), nestin/NES<sup>17,18</sup> in 7/7 ( $28.7 \pm 7.3\%$ ), TIE1 (ref. 10) in 7/7 ( $22.9 \pm 6.2\%$ ), CD144 (VE-cadherin; also known as CDH5)<sup>10</sup> in 5/7 ( $0.5 \pm 0.3\%$ ) and BMPRI1A<sup>19,20</sup> in 7/7 ( $1.5 \pm 0.9\%$ ), and of the stromal marker CD31 (also known as PECAM1)<sup>10</sup> in 6/7 specimens ( $0.7 \pm 0.4\%$ ) (Fig. 1b). Preferential expression by ABCB5<sup>+</sup> compared to ABCB5<sup>−</sup> subpopulations, as previously identified for CD133 (ref. 7), was hereby demonstrated for nestin ( $52.5 \pm 7.9\%$

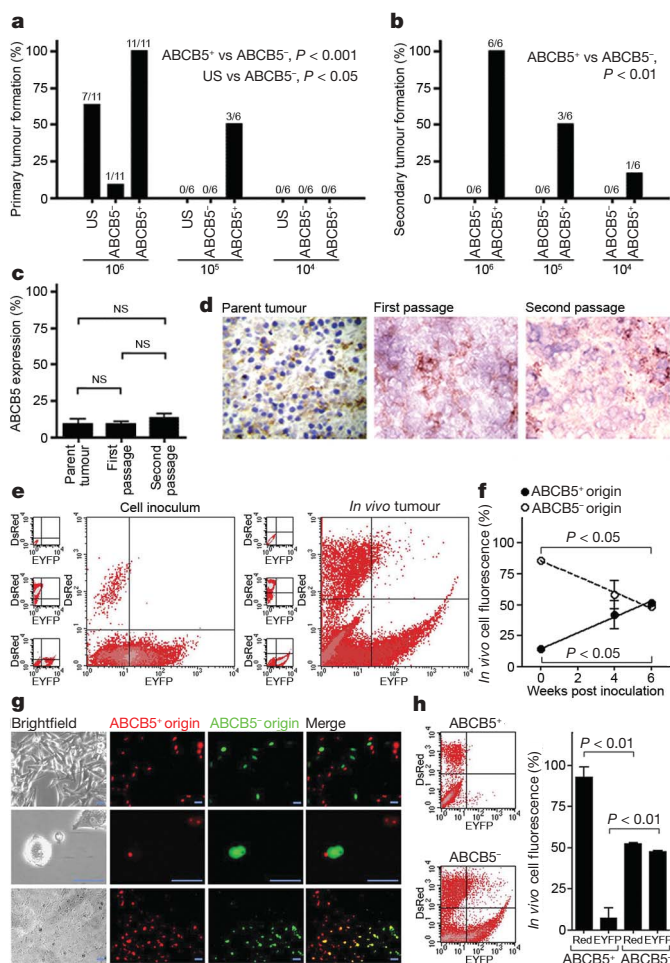


**Figure 1 | ABCB5 expression analyses.** **a**, Melanoma progression tissue microarray analysis for ABCB5, showing significant differences in ABCB5-staining intensities (mean  $\pm$  95% confidence interval (CI); thin or thick nevi versus thin or thick primary melanomas, or versus lymph node or visceral metastases,  $P$  values  $< 0.001$ ; thin versus thick primary melanomas,  $P = 0.004$ ; thin and thick primary melanomas versus lymph node metastases,  $P = 0.001$ ; lymph node versus visceral metastases,  $P = 0.025$ ;  $n$ , provided in figure). The picture colour map corresponds to sample types represented in the core array: green, thin nevi; orange, thick nevi; violet, thin primary melanoma; blue, thick primary melanoma; pink, lymph node metastases; yellow, visceral metastases. The scanning view of ABCB5 staining of the entire array corresponds to the colour key. **b**, Flow cytometry analysis of ABCB5, CD20, nestin, TIE1, CD144, CD31 or BMPRI1A expression in  $n = 7$  melanoma patients. **c**, Marker expression by ABCB5<sup>+</sup> or ABCB5<sup>−</sup> melanoma cells determined by flow cytometry (mean  $\pm$  s.e.m.,  $n = 4–7$  patients).

<sup>1</sup>Transplantation Research Center, Children's Hospital Boston and Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>2</sup>Department of Pathology and <sup>3</sup>Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>4</sup>Department of Surgery, University of Würzburg Medical School, 97080 Würzburg, Germany. <sup>5</sup>Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114, USA. <sup>6</sup>Harvard Skin Disease Research Center, Department of Dermatology, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA.

versus  $24.2 \pm 4.8\%$ , respectively, mean  $\pm$  s.e.m.,  $P = 0.026$ ), TIE1 ( $64.5 \pm 7.6\%$  versus  $22.5 \pm 6.5\%$ ,  $P = 0.002$ ), VE-cadherin ( $12.7 \pm 6.4\%$  versus  $1.0 \pm 0.7\%$ ,  $P = 0.016$ ), and BMPR1A ( $40.9 \pm 6.9$  versus  $2.5 \pm 0.5\%$ ,  $P = 0.001$ ), but not for CD20 ( $0.0 \pm 0.0\%$  versus  $0.8 \pm 0.8\%$ , NS) or CD31 ( $2.4 \pm 1.2\%$  vs.  $0.3 \pm 0.2\%$ , NS) (Fig. 1c). Expression of nestin, TIE1, VE-cadherin and BMPR1A by malignant ABCB5<sup>+</sup> or ABCB5<sup>-</sup> subpopulations within tumours was confirmed by analysis of genetically tracked fluorescent melanoma xenografts (Supplementary Fig. 1). Histologically, ABCB5<sup>+</sup> cells correlated with non-melanized, undifferentiated regions, whereas melanized, more differentiated tumour areas were predominantly ABCB5<sup>-</sup> (Supplementary Fig. 2a).

To determine whether the subset defined by ABCB5 was enriched for MMIC, we compared the abilities of ABCB5<sup>+</sup> versus ABCB5<sup>-</sup> melanoma cells to initiate tumour formation *in vivo*, using



**Figure 2 | Tumorigenicity, self-renewal and differentiation of ABCB5<sup>+</sup> MMIC.** **a**, Primary tumour formation of unsegregated (US), ABCB5<sup>+</sup> or ABCB5<sup>-</sup> cells. **b**, Secondary tumour formation of ABCB5<sup>+</sup> or ABCB5<sup>-</sup> cells. **c**, ABCB5 expression (mean  $\pm$  s.e.m.) in parent tumours ( $n = 3$ ) and respective ABCB5<sup>+</sup>-derived primary ( $n = 11$ ) and secondary ( $n = 7$ ) xenografts. **d**, ABCB5 immunohistochemistry (patient P3). **e–h**, *In vivo* genetic lineage tracking of human ABCB5<sup>+</sup> melanoma cells. **e**, EYFP versus DsRed plots of a genetically labelled inoculum (left) and a corresponding 6-week-old tumour (right). Controls (small panels): non-transfected cells (top), DsRed<sup>+</sup> cells (middle) and EYFP<sup>+</sup> cells (bottom). **f**, Percentage of DsRed<sup>+</sup> or EYFP<sup>+</sup> cells (mean  $\pm$  s.e.m.) in inocula ( $n = 6$ ) and respective tumour xenografts ( $n = 3$ ) as a function of time. **g**, Fluorescence microscopy of dissociated 6-week-old xenografts (top and middle rows) and a corresponding frozen tumour section (bottom row). Scale bars, 25  $\mu$ m. **h**, DsRed/EYFP positivity in ABCB5<sup>+</sup> and ABCB5<sup>-</sup> 6-week-old tumour subpopulations (left); quantified (right) as means  $\pm$  s.d. ( $n = 3$  replicate experiments).

primary-patient-derived tumour cells in serial human-to-NOD/SCID mouse xenotransplantation experiments. ABCB5-dependent cell sorting was performed using immunomagnetic selection<sup>4–7</sup>, followed by confirmation of purity and viability of sorted populations as shown in Supplementary Fig. 3. Groups of mice were transplanted with replicate ( $n = 6–11$ ) inocula of unsegregated, ABCB5<sup>+</sup> or ABCB5<sup>-</sup> melanoma cells representing four distinct patients over a log-fold range from cell doses unable to efficiently initiate tumour growth ( $10^4$  cells) to doses that consistently initiated tumour formation when ABCB5<sup>+</sup> cells were used ( $10^6$  cells) (Fig. 2a, b, and Supplementary Table 1). Of 23 aggregate mice injected with ABCB5<sup>-</sup> melanoma cells, only one transplanted with the highest cell dose generated a tumour. In contrast, 7/23 mice injected with unsegregated populations, and 14/23 mice injected with ABCB5<sup>+</sup> cells formed tumours ( $P < 0.05$  and  $P < 0.001$ , respectively), including all mice injected with the highest cell dose of ABCB5<sup>+</sup> cells (Fig. 2a, and Supplementary Table 1). ABCB5<sup>+</sup> cells re-purified from ABCB5<sup>+</sup>-derived primary xenografts exclusively formed secondary tumours compared to their ABCB5<sup>-</sup> counterparts, in 10/18 versus 0/18 recipients, respectively ( $P < 0.001$ ) (Fig. 2b, and Supplementary Table 1). The MMIC frequency in unsegregated cell populations, calculated as described<sup>5</sup>, was 1/1,090,336 (95% confidence interval, 1/741,780 to 1/1,602,674). The frequencies in ABCB5<sup>+</sup> inocula were 1/158,170 (95% confidence interval, 1/58,464 to 1/427,919) and 1/120,735 (95% confidence interval, 1/44,017 to 1/331,167) for primary and secondary tumour formation, respectively, demonstrating 71-fold and >359-fold enrichment compared to frequencies in ABCB5<sup>-</sup> inocula (1/11,152,529 and <1/43,402,209, respectively). Residual contamination with ABCB5<sup>+</sup> cells (Supplementary Fig. 3a) may account for the single case of tumour formation by an ABCB5<sup>-</sup> inoculum at the highest cell dose and indicates potential underestimation of MMIC enrichment among ABCB5<sup>+</sup> populations. This is suggested by the presence of ABCB5<sup>+</sup> cells in this tumour (Supplementary Fig. 2b) and the concurrent demonstration in genetic lineage tracking experiments that ABCB5<sup>-</sup> melanoma cells do not generate ABCB5<sup>+</sup> progeny (Fig. 2h). Comparison of the cellular diversity of clinical patient tumours with ABCB5<sup>+</sup>-derived primary and secondary xenografts revealed that ABCB5<sup>+</sup> subpopulations re-established parent tumour heterogeneity as determined by flow cytometry (ABCB5 positivity  $9.0 \pm 3.5\%$  (mean  $\pm$  s.e.m.) in parent melanomas and  $8.8 \pm 1.7\%$  and  $13.1 \pm 3.2\%$  in corresponding primary and secondary ABCB5<sup>+</sup>-cell-derived xenografts, respectively) (Fig. 2c, and Supplementary Table 1) or ABCB5 immunohistochemistry (Fig. 2d). Regeneration of patient tumour heterogeneity for ABCB5 and the preferentially co-expressed markers of molecular plasticity and primitive melanoma phenotype CD144 and TIE1 (ref. 10) by primary and secondary ABCB5<sup>+</sup>-cell-derived xenografts was confirmed by immunofluorescent double staining of tumour sections (Supplementary Fig. 2c). In summary, these findings establish that MMIC frequency is markedly enriched in the melanoma minority population defined by ABCB5 and demonstrate *in vivo* self-renewal and differentiation capacity of this subset.

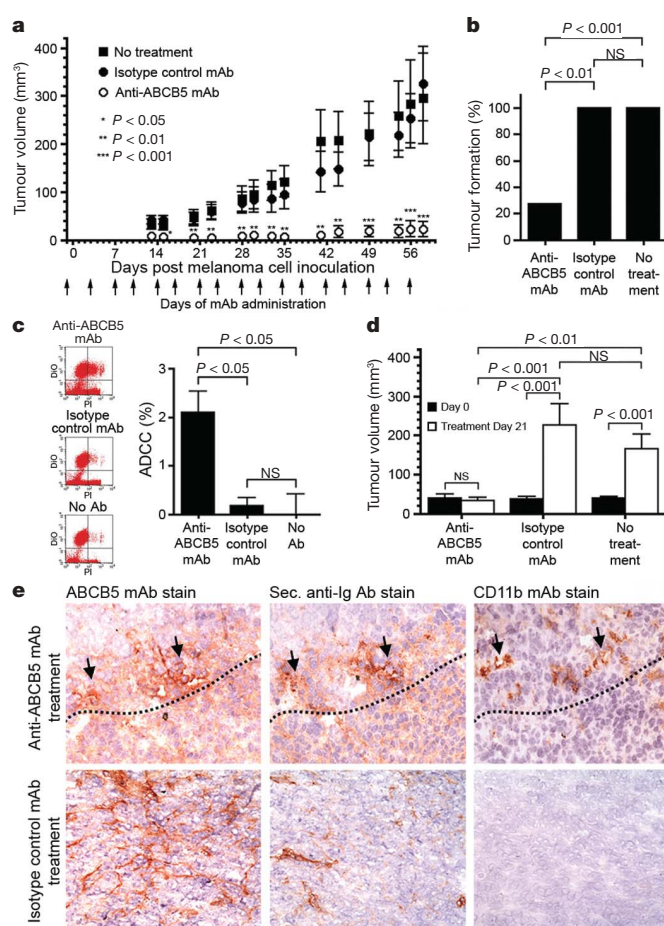
To examine the relative tumour growth contributions of co-xenografted ABCB5<sup>+</sup> and ABCB5<sup>-</sup> subpopulations directly, and to confirm ABCB5<sup>+</sup> self-renewal and differentiation capacity, we isolated ABCB5<sup>+</sup> or ABCB5<sup>-</sup> cells from stably transfected G3361 melanoma cell line variants expressing either red fluorescent protein (DsRed) or enhanced yellow-green fluorescent protein (EYFP), respectively—a model system designed to allow *in vivo* genetic lineage tracking. We found that xenotransplantation of ABCB5<sup>+</sup>/DsRed and ABCB5<sup>-</sup>/EYFP fluorochrome-transfected co-cultures—reconstituted at  $14.0 \pm 3.0\%$  and  $86.0 \pm 3.0\%$  relative abundance (mean  $\pm$  s.d.,  $n = 6$ ), respectively—to NOD/SCID mice resulted in time-dependent, serially increasing relative frequencies of DsRed<sup>+</sup> tumour cells of ABCB5<sup>+</sup> origin in experimental tumours compared to inoculates, up to a frequency of  $51.3 \pm 1.4\%$  at the experimental endpoint of 6 weeks (linear regression slope  $6.4 \pm 1.0$ ,  $P < 0.0001$ )

(Fig. 2e, f, g). These findings establish greater tumorigenicity of ABCB5<sup>+</sup> subsets in a competitive tumour development model. They further indicate that tumour-initiating cells may also drive more differentiated and otherwise non-tumorigenic cancer bulk populations to contribute, albeit less efficiently, to a growing tumour mass. The capacity of non-tumour-initiating cancer cell populations to undergo a limited number of replications is consistent with previous findings in other solid tumours<sup>3–5,21</sup>. Experimental tumours also contained DsRed/EYFP double-positive melanoma cells (Fig. 2e, g), indicating that ABCB5<sup>+</sup>-derived tumour cells, like physiological ABCB5<sup>+</sup> skin progenitors<sup>14</sup>, engage in cell fusion. When ABCB5<sup>+</sup> cells were purified from experimental tumours, we found  $92.9 \pm 6.4\%$  (mean  $\pm$  s.d.,  $n = 3$ ) of fluorescent cells to be of DsRed<sup>+</sup> phenotype (ABCB5<sup>+</sup> origin) (Fig. 2h), confirming self-renewal capacity of this cell subset. EYFP<sup>+</sup>DsRed<sup>−</sup> cells were not found in repeat experiments ( $n = 3$ ) at significant levels among purified ABCB5<sup>+</sup> cells (Fig. 2h), and analysis of ABCB5 expression by triple-colour flow cytometry on DsRed<sup>+</sup> (ABCB5<sup>+</sup> origin) and EYFP<sup>+</sup> (ABCB5<sup>−</sup> origin) subpopulations derived from co-injected *in vivo* tumour xenografts confirmed that ABCB5<sup>+</sup> cells were exclusively of DsRed<sup>+</sup> phenotype, with no significant numbers of EYFP<sup>+</sup>DsRed<sup>−</sup> cells detected (median percentage 0%, results not illustrated). These results indicate that ABCB5<sup>+</sup> tumour cells arose only from ABCB5<sup>+</sup> inocula and that ABCB5<sup>−</sup> cells give rise exclusively to ABCB5<sup>−</sup> progeny. Moreover, fluorescent ABCB5<sup>−</sup> isolates exhibited  $52.5 \pm 0.8\%$  (mean  $\pm$  s.d.,  $n = 3$ ) DsRed positivity (ABCB5<sup>+</sup> origin) and  $47.5 \pm 0.8\%$  EYFP positivity (ABCB5<sup>−</sup> origin) (Fig. 2h), demonstrating that ABCB5<sup>+</sup> melanoma cells possess the capacity to differentiate and give rise to ABCB5<sup>−</sup> tumour populations. These findings confirm the existence of a tumour hierarchy in which ABCB5<sup>+</sup> melanoma cells, enriched for MMIC, self-renew and give rise to more-differentiated ABCB5<sup>−</sup> tumour progeny.

To dissect further and mechanistically whether the ABCB5-defined, MMIC-enriched minority population is also required for tumorigenicity when unsegregated cancer populations are xenografted, we examined whether selective killing of this cell subset can inhibit tumour growth and formation. We administered a monoclonal antibody directed at ABCB5 (refs 7, 14) in a human to the nude mouse melanoma xenograft model, because murine immunoglobulin G1 monoclonal antibodies trigger cellular immune effector functions<sup>22</sup> and because nude as opposed to NOD/SCID mice are capable of tumour cell killing by antibody-dependent cell-mediated cytotoxicity (ADCC)<sup>23</sup>. Anti-ABCB5 monoclonal antibody treatment resulted in significantly inhibited tumour growth compared to that determined in control-monoclonal-antibody-treated or untreated mice over the course of a 58-day observation period (tumour volume for anti-ABCB5-treated ( $n = 11$  mice, no death during the observation period;  $23 \pm 16$  mm<sup>3</sup>; mean  $\pm$  s.e.m.) versus control-monoclonal-antibody-treated ( $n = 10$  mice, excluding 1 death;  $325 \pm 78$  mm<sup>3</sup>),  $P < 0.01$ ; versus untreated ( $n = 18$  mice, excluding 1 death;  $295 \pm 94$  mm<sup>3</sup>),  $P < 0.001$ , see Methods for test used) (Fig. 3a). Control monoclonal antibody treatment showed no significant difference compared to no treatment (Fig. 3a). Anti-ABCB5 monoclonal antibody treatment also significantly inhibited tumour formation, with tumours detected in only 3/11 anti-ABCB5-treated mice versus 10/10 control-antibody-treated mice and 18/18 untreated control animals ( $P < 0.01$  and  $P < 0.001$ , respectively) (Fig. 3b). Human melanoma xenografts grown in untreated nude mice, like those in NOD/SCID recipients, showed tumour heterogeneity for ABCB5 (Supplementary Fig. 4a). Immunohistochemical examination of tumours that successfully grew in the presence of ABCB5 monoclonal antibody revealed that these tumours still contained ABCB5<sup>+</sup> cells (Supplementary Fig. 4b), indicating that ABCB5<sup>+</sup> MMIC had not been fully eradicated. On termination of monoclonal antibody administration, one tumour occurrence was noted among the eight ABCB5-treated mice that had not developed a

tumour during an additional eight-month observation period, indicating prolonged inhibition of tumour-initiating cells.

To determine the mechanism of anti-ABCB5 monoclonal-antibody-mediated inhibition of tumour formation and growth, the immune effector responses ADCC and complement-dependent cytotoxicity (CDC) were assessed, as described<sup>24</sup>. Anti-ABCB5 monoclonal antibody but not isotype control monoclonal antibody significantly induced ADCC-mediated melanoma target cell death ( $2.1 \pm 0.4\%$  versus  $0.2 \pm 0.2\%$ , respectively, mean  $\pm$  s.e.m.,  $P < 0.05$ ) in a melanoma subpopulation comparable in size to the ABCB5-expressing subset<sup>7</sup> (Fig. 3c). Addition of serum to anti-ABCB5-treated cultures in the absence of effector cells, or addition of monoclonal antibody alone did not induce significant cell death compared to controls (results not illustrated), precluding CDC or direct toxic monoclonal antibody effects as significant causes of tumour inhibition.



**Figure 3 | ABCB5 targeting.** **a**, Tumour volumes (mean  $\pm$  s.e.m.) plotted against time. **b**, tumour formation rate in untreated ( $n = 18$ ), control-monoclonal-antibody (mAb)-treated ( $n = 10$ ), or anti-ABCB5 mAb-treated ( $n = 11$ ) animals. **c**, Flow cytometric assessment of ADCC in anti-ABCB5 mAb-treated, control mAb-treated or untreated DiO-labelled melanoma target cultures counterstained with propidium iodide (PI). Left panels, representative flow cytometry results showing lysed, DiO<sup>+</sup>PI<sup>+</sup> target cells in the right upper quadrants. Right panel, analysis of ADCC (mean  $\pm$  s.e.m.) in  $n = 6$  replicate experiments. **d**, Effect of anti-ABCB5 mAb on established melanoma xenografts. Tumour volumes (mean  $\pm$  s.e.m.) for anti-ABCB5 mAb-treated ( $n = 23$ ), control mAb-treated ( $n = 22$ ), or untreated ( $n = 22$ ) animals at days 0 and 21 of treatment. **e**, Immunohistochemistry of patient-derived melanoma xenografts treated with anti-ABCB5 mAb (top) or control mAb (bottom). Adjacent sections were stained with anti-ABCB5 mAb (left), secondary anti-Ig Ab (middle) or CD11b mAb (right), with zones of cellular degeneration in the top row shown below the dotted line.

We next analysed the effects of ABCB5 targeting on established human-to-nude mouse melanoma xenografts ( $n = 13$  derived from three distinct patients and  $n = 10$  derived from established melanoma cultures) to test the hypothesis that negative selection for MMIC by ADCC-mediated ABCB5<sup>+</sup> cell ablation inhibits tumour growth, as would be anticipated in a dynamic *in vivo* situation if the ABCB5<sup>+</sup> melanoma subset is critical to robust tumorigenesis. *In vivo* anti-ABCB5 monoclonal antibody administration, started 14 days following tumour cell inoculation when xenografts were established (day 0), abrogated the significant tumour growth observed in isotype-control-monoclonal-antibody-treated or untreated groups over the course of a 21-day treatment period ( $P < 0.001$  and  $P < 0.001$ , respectively) and significantly inhibited mean tumour volume compared to that determined in either control-treated or untreated mice (tumour volume for anti-ABCB5-treated ( $n = 23$  mice;  $32.7 \pm 9.4 \text{ mm}^3$ ; mean  $\pm$  s.e.m.) versus control-treated ( $n = 22$  mice;  $226.6 \pm 53.8 \text{ mm}^3$ ),  $P < 0.001$ ; versus untreated ( $n = 22$  mice;  $165.4 \pm 36.9 \text{ mm}^3$ ),  $P < 0.01$ , see Methods for test used) (Fig. 3d). The inhibitory effects of ABCB5 monoclonal antibody were also statistically significant when the subsets of freshly patient-derived melanoma xenograft tumours were analysed independently, with abrogation of the significant tumour growth observed in isotype-control-monoclonal-antibody-treated or untreated groups ( $P < 0.05$  and  $P < 0.001$ , respectively) and significantly inhibited mean tumour volume compared to that determined in either control-treated or untreated mice (anti-ABCB5-treated ( $n = 13$  mice;  $29.6 \pm 9.2 \text{ mm}^3$ ) versus control-treated ( $n = 12$  mice;  $289.2 \pm 91.8 \text{ mm}^3$ ),  $P < 0.05$ ; versus untreated ( $n = 12$  mice;  $222.9 \pm 57.5 \text{ mm}^3$ ),  $P < 0.001$ ) (Fig. 3d). Control monoclonal antibody treatment showed no significant effects on tumour growth or tumour volume compared to no treatment in any of the groups analysed. The animals were euthanized following the treatment interval, as required by the applicable experimental animal protocol because of tumour burden and disease state in the patient-derived tumour control groups (measured maximal tumour volume,  $971.5 \text{ mm}^3$ ). Immunohistochemical analysis of anti-ABCB5-treated patient-derived melanoma xenografts revealed only small foci of ABCB5 expression (overall  $< 1\%$  of cells) (focal area of positivity shown in Fig. 3e), corresponding to *in vivo* bound anti-ABCB5 monoclonal antibody in an adjacent section. An additional adjacent section stained for CD11b disclosed macrophage infiltration, corresponding to regions of anti-ABCB5 monoclonal antibody localization, which frequently bordered zones of cellular degeneration and necrosis (Fig. 3e). In contrast, control-treated xenografts revealed 10–15% ABCB5-reactive cells, secondary anti-immunoglobulin monoclonal antibody failed to localize to the respective regions in an adjacent section but detected regions of intravascular murine immunoglobulin, and CD11b<sup>+</sup> macrophages failed to infiltrate the tumour tissue (Fig. 3e). Similar effects were observed in cell-line-derived melanoma xenografts (Supplementary Fig. 4c, d), with enhanced tumour necrosis in anti-ABCB5-treated versus isotype-control-monoclonal-antibody-treated animals (30–40% versus  $< 5\%$  necrotic cells, respectively) (Supplementary Fig. 4c). These findings further support the notion that the ABCB5-defined, MMIC-enriched minority population is required for tumorigenicity.

Because ABCB5 represents a possible chemoresistance mechanism<sup>7,8</sup>, our findings provide evidence for a new, potentially critical link between tumour-initiating cells, cancer progression and chemoresistance in a solid malignancy<sup>25</sup>, raising the possibility that ABCB5<sup>+</sup> MMIC may be responsible both for the progression and chemotherapeutic refractoriness of advanced malignant melanoma, and that MMIC-targeted approaches might therefore ultimately represent novel and translationally relevant therapeutic strategies to disseminated disease. Broader examination of a larger array of clinical specimen is warranted to establish further ABCB5 as a universal MMIC marker and robust candidate therapeutic target. Whether related ABC members<sup>26,27</sup> might also represent prospective markers of tumour-initiating cells, or whether ABCB5 might

represent such a marker in additional malignancies, such as breast cancer, in which it is known to be clinically expressed and specifically downregulated with epigenetic differentiation therapy<sup>28</sup>, requires further study.

Although MMIC are enriched in the melanoma subpopulations defined by ABCB5, clearly not every ABCB5<sup>+</sup> cell represents a MMIC, because purified populations did not invariably form tumours. Our finding that ABCB5 serves as a molecular marker for MMIC is consistent with the demonstration that ABCB5 expression is closely co-regulated with melanotransferrin, a molecule also associated with melanoma growth<sup>29</sup>. The tumour-initiating-cell frequency determined in malignant melanoma is approximately 19-fold lower than that, for example, determined in colon cancer<sup>5</sup>. Tumorigenicity in human-to-mouse xenotransplantation experiments, and as a result calculated stem cell frequency estimates, might vary with the applied experimental conditions, such as the tissue site of xenotransplantation, or the presence or absence of re-activated immune effector mechanisms in recipient immunodeficient mice<sup>30</sup>. Alternatively, inherent differences between stem cell frequencies in distinct malignancies could account for the observed difference. The per cent positivity of tumour cells identified by the prospective marker ABCB5 in clinical melanomas parallels those obtained for the CD133 marker, which detects subpopulations enriched for tumour-initiating cells at similar relative frequencies in brain cancer<sup>4</sup> and colon cancer<sup>5,6</sup>, but likewise does not permit tumour-initiating-cell identification at the clonal level. Further studies are needed to reveal whether tumour-initiating cells can be molecularly defined at the single-cell level in a solid malignancy, or whether more than one cell is necessary for tumour initiation. Our results represent a significant step towards this goal in human malignant melanoma, and provide a basis to elucidate further, and eventually therapeutically target, the specific molecular pathways responsible for tumorigenicity, tumour progression and chemoresistance in tumour-initiating cells.

## METHODS SUMMARY

**Melanocytic tumour progression tissue microarray.** Correlation of ABCB5 expression with melanoma progression was examined using an established microarray<sup>16</sup> and the Chromavision Automated Cellular Imaging System to quantify ABCB5 and control immunostaining intensities.

**Tumour cell isolation and flow cytometry.** Clinical melanoma cells were derived from surgical specimen according to IRB-approved human subjects research protocols. Single-cell suspensions were generated using collagenase. ABCB5 expression was determined by flow cytometry, and ABCB5<sup>+</sup> and ABCB5<sup>−</sup> subpopulations were generated using anti-ABCB5 monoclonal antibody labelling and magnetic-bead cell sorting as described<sup>7,14</sup>. Purity and viability of cell isolates were determined using CD31 and CD45 staining, propidium iodide staining and the calcein-AM assay followed by flow cytometry.

**Human melanoma xenotransplantation and ABCB5 targeting.** NOD/SCID and Balb/c nude mice were maintained under defined conditions in accordance with institutional guidelines and experiments were performed according to approved experimental protocols. For tumorigenicity studies, unsegregated, ABCB5<sup>+</sup>, or ABCB5<sup>−</sup> melanoma cells were injected subcutaneously into flanks of recipient NOD/SCID mice. For MMIC targeting experiments, unsegregated melanoma cells were xenografted subcutaneously into recipient Balb/c nude mice and animals were injected i.p. with anti-ABCB5 monoclonal antibody<sup>7,14</sup> or control monoclonal antibody (500  $\mu\text{g}$  per injection, respectively) bi-weekly starting 24 h before melanoma xenotransplantation or 14 days post tumour cell inoculation, when tumours were established. Tumour formation/growth was assayed as a time course for the duration of the experiment or until excessive tumour burden or disease state required protocol-stipulated euthanasia. ADCC was assessed *in vitro* as described<sup>24</sup> and *in vivo* by histological analysis of tumour-infiltrating immune effector cells.

***In vivo* genetic lineage tracking.** ABCB5<sup>+</sup>/DsRed and ABCB5<sup>−</sup>/EYFP tumour cell populations, immunomagnetically sorted from stably transfected G3361 variants, were reconstituted at desired ratios and injected subcutaneously into recipient NOD/SCID mice. Following xenotransplantation, tumours were serially harvested for determination of relative abundance of DsRed<sup>+</sup> and EYFP<sup>+</sup> melanoma cells.

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Received 13 June; accepted 21 November 2007.**

- Lapidot, T. *et al.* A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645–648 (1994).
- Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nature Med.* **3**, 730–737 (1997).
- Al-Hajj, M. *et al.* Prospective identification of tumorigenic breast cancer cells. *Proc. Natl Acad. Sci. USA* **100**, 3983–3988 (2003).
- Singh, S. K. *et al.* Identification of human brain tumour initiating cells. *Nature* **432**, 396–401 (2004).
- O'Brien, C. A., Pollett, A., Gallinger, S. & Dick, J. E. A human colon cancer cell capable of initiating tumour growth in immunodeficient mice. *Nature* **445**, 106–110 (2007).
- Ricci-Vitiani, L. *et al.* Identification and expansion of human colon-cancer-initiating cells. *Nature* **445**, 111–115 (2007).
- Frank, N. Y. *et al.* ABCB5-mediated doxorubicin transport and chemoresistance in human malignant melanoma. *Cancer Res.* **65**, 4320–4333 (2005).
- Huang, Y. *et al.* Membrane transporters and channels: role of the transportome in cancer chemosensitivity and chemoresistance. *Cancer Res.* **64**, 4294–4301 (2004).
- Chin, L., Garraway, L. A. & Fisher, D. E. Malignant melanoma: genetics and therapeutics in the genomic era. *Genes Dev.* **20**, 2149–2182 (2006).
- Hendrix, M. J., Seftor, E. A., Hess, A. R. & Seftor, R. E. Molecular plasticity of human melanoma cells. *Oncogene* **22**, 3070–3075 (2003).
- Topczewska, J. M. *et al.* Embryonic and tumorigenic pathways converge via Nodal signaling: role in melanoma aggressiveness. *Nature Med.* **12**, 925–932 (2006).
- Fang, D. *et al.* A tumorigenic subpopulation with stem cell properties in melanomas. *Cancer Res.* **65**, 9328–9337 (2005).
- Monzani, E. *et al.* Melanoma contains CD133 and ABCG2 positive cells with enhanced tumorigenic potential. *Eur. J. Cancer* **43**, 935–946 (2007).
- Frank, N. Y. *et al.* Regulation of progenitor cell fusion by ABCB5 P-glycoprotein, a novel human ATP-binding cassette transporter. *J. Biol. Chem.* **278**, 47156–47165 (2003).
- van Kempen, L. C. *et al.* Activated leukocyte cell adhesion molecule/CD166, a marker of tumor progression in primary malignant melanoma of the skin. *Am. J. Pathol.* **156**, 769–774 (2000).
- Kim, M. *et al.* Comparative oncogenomics identifies NEDD9 as a melanoma metastasis gene. *Cell* **125**, 1269–1281 (2006).
- Florenes, V. A. *et al.* Expression of the neuroectodermal intermediate filament nestin in human melanomas. *Cancer Res.* **54**, 354–356 (1994).
- Klein, W. M. *et al.* Increased expression of stem cell markers in malignant melanoma. *Mod. Pathol.* **20**, 102–107 (2007).
- Frank, N. Y. *et al.* Regulation of myogenic progenitor proliferation in human fetal skeletal muscle by BMP4 and its antagonist Gremlin. *J. Cell Biol.* **175**, 99–110 (2006).
- Piccirillo, S. G. *et al.* Bone morphogenetic proteins inhibit the tumorigenic potential of human brain tumour-initiating cells. *Nature* **444**, 761–765 (2006).
- Bao, S. *et al.* Glioma stem cells promote radioresistance by preferential activation of the DNA damage response. *Nature* **444**, 756–760 (2006).
- Hazenbos, W. L. *et al.* Murine IgG1 complexes trigger immune effector functions predominantly via FcγRIII (CD16). *J. Immunol.* **161**, 3026–3032 (1998).
- Kanazawa, J. *et al.* Therapeutic potential of chimeric anti-(ganglioside GD3) antibody KM871: antitumor activity in xenograft model of melanoma and effector function analysis. *Cancer Immunol. Immunother.* **49**, 253–258 (2000).
- Kroesen, B. J. *et al.* Direct visualisation and quantification of cellular cytotoxicity using two colour fluorescence. *J. Immunol. Methods* **156**, 47–54 (1992).
- Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105–111 (2001).
- Dean, M., Fojo, T. & Bates, S. Tumour stem cells and drug resistance. *Nature Rev. Cancer* **5**, 275–284 (2005).
- Patrawala, L. *et al.* Side population is enriched in tumorigenic, stem-like cancer cells, whereas ABCG2<sup>+</sup> and ABCG2<sup>−</sup> cancer cells are similarly tumorigenic. *Cancer Res.* **65**, 6207–6219 (2005).
- Arce, C. *et al.* A proof-of-principle study of epigenetic therapy added to neoadjuvant Doxorubicin cyclophosphamide for locally advanced breast cancer. *PLoS ONE* **1**, e98 (2006).
- Suryo Rahmanto, Y., Dunn, L. & Richardson, D. Identification of distinct changes in gene expression after modulation of melanoma tumor antigen p97 (melanotransferrin) in multiple models *in vitro* and *in vivo*. *Carcinogenesis* **28**, 2172–2183 (2007).
- Kelly, P. N. *et al.* Tumor growth need not be driven by rare cancer stem cells. *Science* **317**, 337 (2007).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank D. Herlyn and M. Herlyn for providing fresh melanoma tissue specimen for our studies. The construction of the tissue microarray was possible only through the collaborative assistance of P. Van Belle, D. Elder, V. Prieto and A. Lazar. The tissue microarrays were performed with the technical assistance of R. Kim, K. Lamb and L. Biagini. We thank A. Baldor for technical assistance with tumour xenotransplantation experiments, and M. Grimm for tissue sectioning and immunohistochemistry. We thank D. Scadden for comments on the manuscript. This work was supported by the NCI/NIH (M.H.F.), a NCI/NIH Specialized Program of Research Excellence (SPORE) in Skin Cancer (T.S.K.) and the Department of Defense (M.H.F.).

**Author Contributions** T.S., N.Y.F., and M.H.F. planned the project. T.S., N.Y.F., K.Y., A.M.W.-G., Q.Z., S.J. and C.W. carried out experimental work. T.S., G.F.M., N.Y.F., A.M.W.-G., R.C.F. T.S.K., M.H.S. and M.H.F. analysed data. G.F.M., Q.Z., A.M.W.-G., M.G. and L.M.D. provided clinical information and human tissues or performed pathological analysis. T.S., G.F.M., N.Y.F. and M.H.F. wrote the paper. All authors discussed the results and commented on the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details accompany the HTML version of the paper at [www.nature.com/nature](http://www.nature.com/nature). Correspondence and requests for materials should be addressed to M.H.F. ([mfrank@rics.bwh.harvard.edu](mailto:mfrank@rics.bwh.harvard.edu)).

## METHODS

**Melanoma cells and culture methods.** The ABCB5-expressing G3361 human malignant melanoma cell line<sup>7,14</sup>, derived from a single tumour cell cloned in soft agar, was provided by E. Frei III and cultured as previously described<sup>7</sup>. The G3361/DsRed and G3361/EYFP cell lines were generated by stable transfection of G3361 melanoma cells with either *Discosoma* sp. red fluorescent protein (DsRed) or the enhanced yellow-green variant (EYFP) of the *Aequorea victoria* green fluorescent protein (GFP) in conjunction with the simian virus 40 large T-antigen nuclear retention signal, using pDsRed-Nuc or pEYFP-Nuc mammalian expression vectors also containing a neomycin resistance cassette (BD Biosciences) and the Lipofectamine 2000 reagent (Invitrogen), as previously described<sup>14</sup>. Clonal G3361/DsRed and G3361/EYFP cultures were generated from stably transfected cultures by limiting dilution. Clinical melanoma cells ( $n = 7$  patients) were freshly derived from surgical specimens according to human subjects research protocols approved by the IRBs of the University of Würzburg Medical School or the Wistar Institute.

**Antibodies.** The specific IgG1 $\kappa$  anti-ABCB5 monoclonal antibody (mAb) 3C2-1D12 (refs 7, 14) was used in the herein reported studies. Unconjugated or FITC-conjugated MOPC-31C mouse isotype control mAbs, FITC-conjugated goat anti-mouse IgG secondary Ab, phycoerythrin (PE)-conjugated anti-human CD20, anti-human and anti-mouse CD31, anti-human and anti-mouse CD45, and isotype control mAbs were purchased from Pharmingen. Allophycocyanin (APC)-conjugated and PE-conjugated secondary mAbs were purchased from eBioscience. Unconjugated anti-human TIE1, anti-human BMPRIA, PE-conjugated anti-human VE-cadherin and anti-human nestin mAbs were from R&D Systems. The following antibodies were used for immunohistochemistry or immunofluorescence staining: mouse anti-ABCB5 mAb<sup>7,14</sup>, HRP-conjugated horse anti-mouse IgG secondary Ab, HRP-conjugated horse anti-goat IgG secondary Ab and HRP-conjugated goat anti-rabbit IgG secondary Ab (Vector Laboratories), FITC-conjugated rabbit anti-mouse IgG secondary Ab (ZYMED Laboratories), unconjugated rabbit anti-human VE-cadherin Ab (provided by Cell Signaling Technology), mouse control IgG Abs (DAKO), goat anti-human TIE1 Ab (Neuromics), rat anti-mouse CD11b Ab and rat anti-mouse CD31 Ab (BD Biosciences Pharmingen), rabbit anti-human CD31 Ab (Bethyl Laboratories), donkey anti-mouse IgG-AF488, donkey anti-rabbit IgG-AF594, donkey anti-rat IgG-AF594 and donkey anti-goat IgG-AF594 (Invitrogen), Texas Red-conjugated donkey anti-rabbit IgG secondary Ab, and rabbit control IgG Ab (all from Jackson ImmunoResearch).

**Histopathology and immunohistochemistry.** Five micron-thick melanoma cryosections were fixed in  $-20^{\circ}\text{C}$  acetone for 5 min. Air-dried sections were incubated with  $10\text{ }\mu\text{g ml}^{-1}$  ABCB5 mAb or  $2.5\text{ }\mu\text{g ml}^{-1}$  CD11b mAb at  $4^{\circ}\text{C}$  overnight;  $10$  or  $2.5\text{ }\mu\text{g ml}^{-1}$  mouse IgG were used as negative controls. Sections were washed with PBS 3 times for 5 min each and incubated with 1:200 peroxidase-conjugated secondary Abs for ABCB5 or CD11b staining. For ABCB5/VE-cadherin, ABCB5/TIE1, or ABCB5/CD31 fluorescence double labelling,  $5\text{ }\mu\text{m}$  melanoma sections were fixed in  $-20^{\circ}\text{C}$  acetone for 5 min. Air-dried sections were incubated with  $10\text{ }\mu\text{g ml}^{-1}$  ABCB5 mAb and  $2.5\text{ }\mu\text{g ml}^{-1}$  VE-cadherin, TIE1 or CD31 Abs at  $4^{\circ}\text{C}$  overnight;  $10\text{ }\mu\text{g ml}^{-1}$  mouse IgG and  $2.5\text{ }\mu\text{g ml}^{-1}$  rabbit IgG were used as negative controls. Sections were washed three times with PBS containing 0.05% Tween 20 for 5 min each and incubated with a 1:150 dilution of Texas Red-conjugated or AF594-conjugated secondary Abs and FITC-conjugated rabbit anti-mouse IgG Ab for 30 min at room temperature. After subsequent washings, the sections were mounted with VECTASHIELD mounting medium (Vector Laboratories) and covered with a cover slip. Immunofluorescence reactivity was viewed on an Olympus BX51/52 system microscope coupled to a Cytovision system (Applied Imaging).

**Tissue microarray design and analysis.** The Melanocytic Tumour Progression tissue microarray (TMA) is the product of a joint effort of three Skin SPORs (Harvard Medical School, MD Anderson Cancer Center, University of Pennsylvania). This array contains  $480 \times 0.6\text{ mm}$  cores of tumour tissue representing four major diagnostic tumour types: benign nevi, primary cutaneous melanoma, lymph node metastasis and visceral metastasis. Cases were collected from the Pathology services of the three participating institutions. For quality control purposes, two duplicate cores are chosen at each distinct region. Nevi and primary melanomas had either one region or three regions of the tissue block sampled (2 or 6 cores), whereas metastatic tumours had one region sampled from each block. Therefore, the 480 cores represent 2 adjacent cores from 240 distinct histological regions. This array includes 130 cores from 35 nevi, 200 cores from 60 primary melanoma and 150 cores from 75 metastatic lesions. Operationally, thin nevi and thin melanomas involved only the superficial/papillary dermis, whereas thick nevi and thick melanomas had grown to involve both papillary and deep (reticular) dermis. This array was constructed in the laboratory of M. Rubin. Histological sections of the tissue array slide were baked

at  $58^{\circ}\text{C}$  for 20 min and then treated with the following: xylene (twice for 1 h, then 10 min), 100% ethanol twice for 2 min, 95% ethanol for 2 min, and  $\text{dH}_2\text{O}$  three times for 2 min. Antigen retrieval was performed in  $10\text{ mmol l}^{-1}$  citrate buffer, pH 6.0 with boiling in a pressure cooker for 10 min and then cooling to room temperature. After washing with PBS twice for 5 min, tissue was blocked with 10% horse serum and 1% BSA in PBS at room temperature for 1 h then incubated with  $5\text{ }\mu\text{g ml}^{-1}$  ABCB5 mAb at  $4^{\circ}\text{C}$  overnight. The tissue was then washed three times with PBS-0.05% Tween 20 for 5 min then treated with 3%  $\text{H}_2\text{O}_2$ /PBS for 15 min. After rinsing in PBS, the sections were incubated with 1:200 biotinylated horse anti-mouse IgG Ab at room temperature for 30 min, rinsed in PBS-Tween three times for 5 min, and incubated with avidin-biotin-horseradish peroxidase complex (Vector Laboratories) for 30 min at room temperature. Immunoreactivity was detected using NovaRed substrate (Vector Laboratories). The Chromavision Automated Cellular Imaging System (ACIS) was used to quantify the immunostaining intensity of ABCB5 and mIgGIR on the HTMA 84 tissue microarray. The control slide intensity values (background plus intrinsic melanization) were subtracted from the experimental slide and the difference in the intensity values for each core was taken to be the true staining. The graph in Fig. 1a shows with 95% confidence interval the difference in intensity for each pathology diagnosis.  $P$  values between relevant groups were calculated using the independent/samples  $t$ -test. The number above each error bar shows the number of cases within each group.

**Flow cytometric analysis.** Analysis of ABCB5, CD20, CD31, CD45, VE-cadherin, BMPRIA, nestin, or TIE1 expression, or of co-expression of ABCB5 with the CD20, CD31, VE-cadherin or BMPRIA surface markers or the nestin or TIE1 intracellular markers in clinical patient-derived melanoma cell suspensions or in G3361 melanoma cells was performed by single- or dual-colour flow cytometry, as described previously<sup>7</sup>. Co-expression analyses of ABCB5 with the above-listed markers in single-cell suspensions derived from G3361/EYFP tumour xenografts and expression analyses of ABCB5 in G3361/DsRed-G3361/EYFP-derived tumours were performed by triple-colour flow cytometry, gating on EYFP-expressing melanoma cells or ABCB5-expressing cells, respectively. Clinical melanoma cells were incubated with anti-ABCB5 mAb or isotype control mAb or no Ab followed by counterstaining with APC-conjugated donkey anti-mouse IgG. Cells were then fixed in PBS containing 2% paraformaldehyde (30 min at  $4^{\circ}\text{C}$ ), and subsequently incubated with PE-conjugated anti-CD20, anti-CD31, anti-VE-cadherin, anti-nestin or PE-conjugated isotype control mAbs, or unconjugated anti-BMPRIA, anti-TIE1 or unconjugated isotype control mAbs followed by counterstaining with PE- or FITC-conjugated anti-immunoglobulin secondary antibodies. Washing steps with staining buffer or 1% saponin permeabilization buffer were performed between each step. Dual- or triple-colour flow cytometry was subsequently done with acquisition of fluorescence emission at the FL1 (FITC, EYFP) and/or FL2 (PE, DsRed) and FL4 (APC) spectra on a Becton Dickinson FACScan (Becton Dickinson), as described<sup>7</sup>. Statistical differences between expression levels of the above-listed markers by ABCB5<sup>+</sup> and ABCB5<sup>-</sup> patient-derived melanoma cells were determined using the nonparametric Mann-Whitney test. A two-sided  $P$  value of  $P < 0.05$  was considered significant.

**Cell isolation.** Single-cell suspensions were generated from human melanoma xenografts on surgical dissection of tumours from euthanized mice. Each tumour was cut into small pieces ( $\sim 1\text{ mm}^3$ ) and tumour fragments were subsequently incubated in 10 ml sterile PBS containing  $0.1\text{ g l}^{-1}$  calcium chloride and  $5\text{ }\mu\text{g ml}^{-1}$  Collagenase Serva NB6 (SERVA Electrophoresis GmbH) for 3 h at  $37^{\circ}\text{C}$  on a shaking platform at 200 r.p.m. to generate single-cell suspensions. Subsequently, tumour cells were washed with PBS for excess collagenase removal. ABCB5<sup>+</sup>-purified (ABCB5<sup>+</sup>) cells were isolated by positive selection and ABCB5<sup>+</sup>-depleted (ABCB5<sup>-</sup>) cell populations were generated by removing ABCB5<sup>+</sup> cells using anti-ABCB5 mAb labelling and magnetic-bead cell sorting as described<sup>4,7,14</sup>. Briefly, human G3361 melanoma cells or single-cell suspensions derived from human melanoma xenografts or clinical melanoma samples were labelled with anti-ABCB5 mAb ( $20\text{ }\mu\text{g ml}^{-1}$ ) for 30 min at  $4^{\circ}\text{C}$ , washed for excess antibody removal, followed by incubation with secondary anti-mouse IgG mAb-coated magnetic microbeads (Miltenyi Biotec) and subsequent dual-passage cell separation in MiniMACS separation columns (Miltenyi Biotec), according to the manufacturers recommendations. Purity of ABCB5<sup>+</sup> and ABCB5<sup>-</sup> (ABCB5<sup>+</sup> cell-depleted) clinical melanoma cell isolates or of ABCB5<sup>+</sup> and ABCB5<sup>-</sup> cell isolates derived from ABCB5<sup>+</sup> patient cell-derived primary melanoma xenograft cells was assayed following magnetic-bead cell sorting by incubation with FITC-conjugated goat anti-mouse IgG secondary Ab and subsequent flow cytometric analysis of ABCB5 expression. ABCB5<sup>+</sup> cell purification resulted in 10.4-fold enrichment of ABCB5<sup>+</sup> melanoma cell frequency from  $8.9 \pm 1.4\%$  in unsegregated samples to  $92.4 \pm 2.8\%$  (mean  $\pm$  s.e.m.,  $P < 0.001$ , Supplementary Fig. 2a). Negative selection for ABCB5<sup>+</sup> cells resulted in 6.7-fold depletion of ABCB5<sup>+</sup> cell frequency to  $1.3 \pm 0.6\%$  (mean  $\pm$

s.e.m.,  $P < 0.05$ , Supplementary Fig. 2a). CD31<sup>+</sup> or CD45<sup>+</sup> cell frequencies among unsegregated, ABCB5<sup>+</sup> or ABCB5<sup>-</sup> cell suspensions were determined by single colour flow cytometry, as above. Statistical differences in marker expression between unsegregated, ABCB5<sup>+</sup>, and ABCB5<sup>-</sup> human melanoma cells were determined using parametric ANOVA or the nonparametric Kruskal–Wallis Test followed by Dun's correction for comparisons of multiple groups, with two-tailed  $P$  values  $< 0.05$  considered significant.

**Animals.** Balb/c nude mice and NOD/SCID mice were purchased from The Jackson Laboratory. Mice were maintained in accordance with the institutional guidelines of Children's Hospital Boston and Harvard Medical School and experiments were performed according to approved experimental protocols.

**Human melanoma xenotransplantation.** Unsegregated, ABCB5<sup>+</sup>, or ABCB5<sup>-</sup> clinical patient-derived melanoma cells ( $10^6$ ,  $10^5$  or  $10^4$  per inoculum), or ABCB5<sup>+</sup> or ABCB5<sup>-</sup> cells isolated from primary ABCB5<sup>+</sup> patient-derived xenografts ( $10^6$ ,  $10^5$ , or  $10^4$  per inoculum) were injected subcutaneously uni- or bilaterally into the flanks of recipient NOD/SCID mice. Tumour formation/growth was assayed weekly as a time course, at least up to the endpoint of 8 weeks, unless excessive tumour size or disease state required protocol-stipulated euthanasia earlier, by determination of tumour volume (TV) according to the established formula  $[TV \text{ (mm}^3\text{)} = \pi / 6 \times 0.5 \times \text{length} \times (\text{width})^2]$ . With respect to tumour formation, mice were considered tumour-negative if no tumour tissue was identified on necropsy. Statistically significant differences in primary and secondary tumour formation were assessed using the Fisher's Exact test. Differences in tumour volumes were determined using one-way ANOVA followed by the Bonferroni correction or the Kruskal–Wallis Test followed by Dun's correction, with two-tailed  $P$  values  $< 0.05$  considered significant. Tumour-initiating cell frequencies and respective confidence intervals were calculated as previously described<sup>5</sup>, using the L-Calc version 1.1 statistical software program for limiting dilution analysis (Stemcell Technologies).

**In vivo genetic lineage tracking.** ABCB5<sup>+</sup>/DsRed and ABCB5<sup>-</sup>/EYFP human G3361 tumour cell populations, generated using magnetic-bead cell sorting as above, were reconstituted at the desired ratios on the basis of cell counting and the resultant relative abundance ratios in inocula were determined by dual-colour flow cytometry (F11 (EYFP) versus F12 (DsRed) plots) before xenotransplantation. G3361/DsRed and G3361/EYFP co-cultures were injected subcutaneously ( $10^7$  cells per inoculum) into the right flank of recipient NOD/SCID mice. At 4 or 6 weeks post xenotransplantation, tumours were harvested and single-cell suspensions or frozen tissue sections prepared as above, for determination of relative *in vivo* abundance of DsRed<sup>+</sup> and EYFP<sup>+</sup> melanoma cells by dual-colour flow cytometry or fluorescence microscopy of tumour-derived single-cell suspensions (on attachment in adherent tissue culture plates), and for analysis of 5  $\mu\text{m}$  frozen tissue sections by fluorescence microscopy. Percentages were calculated as follows: %DsRed<sup>+</sup> cells = (%DsRed<sup>+</sup> / (%DsRed<sup>+</sup> + %EYFP<sup>+</sup>)  $\times 100$ ) and %EYFP<sup>+</sup> cells = (%EYFP<sup>+</sup> / (%DsRed<sup>+</sup> + %EYFP<sup>+</sup>)  $\times 100$ ). In additional experiments, the relative abundance of DsRed<sup>+</sup> and EYFP<sup>+</sup> melanoma cells was determined by dual-colour flow cytometry as above in ABCB5<sup>+</sup> or ABCB5<sup>-</sup> subsets purified from xenografts, or by triple-colour flow cytometry of unsorted, freshly dissociated xenografts gating on ABCB5-expressing cells (APC, F14 fluorescence), and the percentages of DsRed<sup>+</sup> and EYFP<sup>+</sup> tumour cells were statistically compared using the unpaired Student's  $t$ -test, with a two-sided  $P$  value of  $P < 0.05$  considered statistically significant. FACS-sorting of tumour xenograft cells of ABCB5<sup>+</sup>/DsRed (F12 fluorescence) versus ABCB5<sup>-</sup>/EYFP (F11 fluorescence) origin for real-time RT–PCR analysis of BMPR1A, VE-cadherin, nestin and TIE1 expression was performed on a dual-laser FACSVantage flow cytometer (Becton Dickinson). Flow cytometric co-expression analysis of ABCB5 with the CD20, CD31, VE-cadherin, BMPR1A, nestin, or TIE1 markers was performed on single tumour cell suspensions prepared from xenograft tumours induced by inoculation of unsegregated G3361/EYFP tumour cells ( $10^7$  cells per inoculum) into recipient NOD/SCID mice.

**Anti-ABCB5 mAb targeting.** For targeting experiments directed at tumour formation, unsegregated human G3361 melanoma cells were xenografted subcutaneously into recipient Balb/c nude mice ( $10^7$  per inoculum). Animals were injected intraperitoneally with anti-ABCB5 mAb (clone 3C2-1D12)<sup>7,14</sup> (500  $\mu\text{g}$  per injection) or isotype control mAb (500  $\mu\text{g}$  per injection) bi-weekly, or no Ab starting 24 h before melanoma xenotransplantation. Tumour growth was assayed bi-weekly as a time course by determination of tumour volume, as described above. For targeting experiments directed at established melanoma xenografts, unsegregated primary-patient-derived or human G3361 melanoma cells were xenografted subcutaneously into the right flank of recipient Balb/c nude mice ( $10^7$  per inoculum). Fourteen days post tumour cell inoculation (day 0), tumour volumes were determined, and mice were randomized into three treatment groups (anti-ABCB5 mAb treatment, isotype control mAb treatment or no treatment), with groups consisting of  $n = 21$ –22 animals, comprising  $n = 12$ –13 mice bearing primary-patient-derived tumours ( $n = 5$  derived from

patient P1,  $n = 3$  from patient P3,  $n = 4$ –5 from patient P7, Supplementary Table 1) and  $n = 10$  mice bearing human-cell-line-derived tumours. Tumour volumes at day 0 did not significantly differ among the groups ( $39.8 \pm 9.3$  versus  $37.5 \pm 6.7$  versus  $38.2 \pm 5.9 \text{ mm}^3$ , respectively, mean  $\pm$  s.e.m., NS), and furthermore did not significantly differ among the subgroups of primary patient-derived tumours ( $48.5 \pm 15.7$  versus  $44.4 \pm 11.1$  versus  $45.7 \pm 9.4 \text{ mm}^3$ , respectively, mean  $\pm$  s.e.m., NS) or cell-line-derived tumours ( $28.4 \pm 5.8$  versus  $29.3 \pm 6.1$  versus  $29.1 \pm 6.0 \text{ mm}^3$ , respectively, mean  $\pm$  s.e.m., NS). Subsequently, mice were injected intraperitoneally with anti-ABCB5 mAb (clone 3C2-1D12)<sup>7,14</sup> (500  $\mu\text{g}$  per injection) or isotype control mAb (500  $\mu\text{g}$  per injection) or no Ab bi-weekly for the duration of the experiment. Tumour formation/growth was assayed weekly as a time course by determination of tumour volume as described above, until excessive tumour burden or disease state required protocol-stipulated euthanasia. Differences in tumour volumes were determined using parametric ANOVA or the non-parametric Kruskal–Wallis Test followed by Dun's correction for comparisons of multiple groups, with two-tailed  $P$  values  $< 0.05$  considered significant. Differences in tumour volumes at different time points within experimental groups were determined using parametric ANOVA (repeated measures (paired) test) or the non-parametric Kruskal–Wallis Test (repeated measures (paired) test).

**Assessment of ADCC and CDC.** ADCC or CDC was determined by the established method of dual-colour flow cytometry. Briefly, human G3361 melanoma cell suspensions in serum-free Dulbecco's Modified Eagle's Medium (DMEM) (BioWhittaker) were labelled with 3,3'-diiodoacetylcarboxycyanine (DiO) (Invitrogen) according to the manufacturer's recommendations. DiO-labelled melanoma cells were then plated at a density of  $3 \times 10^5$  cells per well in flat-bottomed 6-well culture plates in 3 ml and cultured in standard medium in a humidified incubator overnight. Thereafter, DiO-labelled melanoma target cells were pre-incubated in the presence or absence of anti-ABCB5 or isotype control mAbs (20  $\mu\text{g ml}^{-1}$ , respectively) for 30 min at  $37^\circ\text{C}$ , 5%  $\text{CO}_2$ , and subsequently co-cultured for additional 24 h at  $37^\circ\text{C}$ , 5%  $\text{CO}_2$  with or without freshly isolated Balb/c nude mouse effector splenocytes ( $12 \times 10^6$  cells per well, 1:40 target to effector cell ratio) for assessment of ADCC, or in the presence or absence of 5% Balb/c nude mouse serum for determination of CDC. Subsequently, cells and their supernatants were harvested and analysed by dual-colour flow cytometry on a FACSCalibur machine (Becton Dickinson) immediately on addition of 10  $\mu\text{g ml}^{-1}$  propidium iodide (PI) (Sigma), with lysed target cells recognized by a DiO<sup>+</sup>PI<sup>+</sup> phenotype. ADCC levels for the three treatment groups were calculated as follows: [ADCC (%) = (DiO<sup>+</sup>PI<sup>+</sup> % sample positivity) – (mean Ab-untreated DiO<sup>+</sup>PI<sup>+</sup> % sample positivity)]. Differences in ADCC levels were determined using non-parametric one-way ANOVA (Kruskal–Wallis Test) followed by Dun's correction, with two-tailed  $P$  values  $< 0.05$  considered significant.

**Cell viability measurements.** Cell viability was measured in tumour cell inocula before xenotransplantation using calcein-AM staining. Briefly,  $1 \times 10^6$  unsegregated, ABCB5<sup>+</sup>, or ABCB5<sup>-</sup> melanoma cells were incubated with calcein-AM (Molecular Probes) for 30 min at  $37^\circ\text{C}$  and 5%  $\text{CO}_2$  to allow for substrate uptake and enzymatic activation to the fluorescent derivative. Subsequently the cells were washed and fluorescence measurements acquired by flow cytometry at the F12 emission spectrum on a Becton Dickinson FACScan. Cells exhibiting generation of the fluorescent calcein-AM derivative compared to unexposed samples were considered viable. Cell viability was also determined using the trypan blue dye exclusion method.

**RNA extraction and real-time quantitative reverse transcription PCR (RT–PCR).** Real-time RT–PCR for BMPR1A, VE-cadherin, nestin and TIE1 gene expression analysis were performed as follows: total RNA was extracted from melanoma cells using the RNeasy Micro kit (Qiagen). Total RNA (5  $\mu\text{g}$ ) in 20  $\mu\text{l}$  RT reaction mix was transcribed into complementary DNA using the SuperScript III First-Strand Synthesis System for RT–PCR (Invitrogen). All reagents for real-time RT–PCR were from Applied Biosystems. The assay numbers for human  $\beta$ -actin, BMPR1A, VE-cadherin, nestin and TIE1 were 4310881E, Hs01034909\_g1, Hs00174344\_ml, Hs00707120\_sl and Hs00178500\_ml, respectively. Real-time quantitative RT–PCR was performed on a 7300 real-time PCR System (Applied Biosystems) in a 25  $\mu\text{l}$  reaction mix containing 1  $\mu\text{l}$  cDNA,  $1 \times$  TaqMan Universal PCR Master Mix and  $1 \times$  of each of the assays. Thermocycling was carried out at  $50^\circ\text{C}$  for 2 min,  $95^\circ\text{C}$  for 10 min, followed by 40 cycles at  $95^\circ\text{C}$  for 15 s and  $60^\circ\text{C}$  for 1 min. All samples were run in triplicate. The relative amounts of BMPR1A, VE-cadherin, nestin and TIE1 transcripts were analysed using the  $2^{-\Delta\Delta C_T}$  method, as described previously<sup>19</sup>. Statistical differences between messenger RNA expression levels of the above-listed markers by fluorescent xenograft cells of ABCB5<sup>+</sup>/DsRed origin and ABCB5<sup>-</sup>/EYFP origin were determined using the non-parametric Student's  $t$ -test. A two-sided  $P$  value of  $P < 0.05$  was considered significant.

**Quantification of DNA content by propidium iodide (PI) staining.** Freshly sorted ABCB5<sup>+</sup> or ABCB5<sup>+</sup>-depleted (ABCB5<sup>-</sup>) clinical melanoma cells or ABCB5<sup>+</sup> patient cell-derived primary melanoma xenograft cells were fixed in ice-cold 65% (v/v) ethanol in PBS, washed in cold PBS, and incubated in a PI-staining mixture followed by determination of the cell fraction containing < 2*n* DNA by flow cytometry (Becton Dickinson FACSscan), as described previously<sup>7,14</sup>. The frequency of cellular fragments, non-viable cells and/or contaminating blood components containing < 2*n* DNA comprised  $2.8 \pm 0.3\%$  versus  $2.7 \pm 1.9\%$  (mean  $\pm$  s.e.m.) in patient-derived ABCB5<sup>+</sup> or ABCB5<sup>-</sup> cell suspensions, respectively, and  $1.4 \pm 0.2\%$  versus  $0.6 \pm 0.1\%$  (mean  $\pm$  s.e.m.) in ABCB5<sup>+</sup> and ABCB5<sup>-</sup> cell isolates derived from ABCB5<sup>+</sup> patient cell-derived primary melanoma xenograft cells, respectively, with no significant differences detected among isolates, when subjected to the non-parametric Mann–Whitney test (Supplementary Fig. 2d).

## LETTERS

# Listeriolysin O allows *Listeria monocytogenes* replication in macrophage vacuoles

Cheryl L. Birmingham<sup>1,2</sup>, Veronica Canadien<sup>1</sup>, Natalia A. Kaniuk<sup>1</sup>, Benjamin E. Steinberg<sup>1,3</sup>, Darren E. Higgins<sup>4</sup> & John H. Brummell<sup>1,2,3</sup>

*Listeria monocytogenes* is an intracellular bacterial pathogen that replicates rapidly in the cytosol of host cells during acute infection<sup>1</sup>. Surprisingly, these bacteria were found to occupy vacuoles in liver granuloma macrophages during persistent infection of severe combined immunodeficient (SCID) mice<sup>2</sup>. Here we show that *L. monocytogenes* can replicate in vacuoles within macrophages. In livers of SCID mice infected for 21 days, we observed bacteria in large LAMP1<sup>+</sup> compartments that we termed spacious *Listeria*-containing phagosomes (SLAPs). SLAPs were also observed *in vitro*, and were found to be non-acidic and non-degradative compartments that are generated in an autophagy-dependent manner. The replication rate of bacteria in SLAPs was found to be reduced compared to the rate of those in the cytosol. Listeriolysin O (LLO, encoded by *hly*), a pore-forming toxin essential for *L. monocytogenes* virulence<sup>1</sup>, was necessary and sufficient for SLAP formation. A *L. monocytogenes* mutant with low LLO expression was impaired for phagosome escape but replicated slowly in SLAPs over a 72 h period. Therefore, our studies reveal a role for LLO in promoting *L. monocytogenes* replication in vacuoles and suggest a mechanism by which this pathogen can establish persistent infection in host macrophages.

*L. monocytogenes* is a Gram-positive bacterial pathogen that causes acute infection in immunocompromised individuals and pregnant women<sup>1</sup>. After entry into host cells, this pathogen initially occupies a phagosome. LLO, a cholesterol-dependent pore-forming toxin<sup>3</sup>, blocks phagosome-lysosome fusion by generating small pores that uncouple pH and calcium gradients across the phagosome membrane<sup>4</sup>. A second function for LLO, in concert with the action of two phospholipases, is to promote phagosome escape by the bacteria<sup>1</sup>. Once within the cytosol, *L. monocytogenes* replicates rapidly and usurps the host actin polymerization machinery to move through the cytosol and spread into neighbouring cells<sup>1</sup>. LLO is essential for virulence in animal models of infection<sup>1</sup> and its function is known to be impaired by host innate immune defences<sup>5,6</sup>. LLO is also a major antigen for adaptive immune responses, which normally mediate clearance of *L. monocytogenes* infection<sup>7</sup>.

In severe combined immunodeficient (SCID) mice, which lack adaptive immunity, *L. monocytogenes* can cause persistent infection<sup>2</sup>. In these mice, bacteria are localized to macrophages in tissue granulomas (particularly within the liver) and are largely absent from other cell types<sup>2</sup>. Surprisingly, *L. monocytogenes* occupy vacuoles during persistent infection, although the nature of these compartments is unclear (Fig. 1a)<sup>2</sup>. To characterize *Listeria*-containing vacuoles in host cells during persistent infection, we analysed liver sections from SCID mice that had been infected with wild-type *L. monocytogenes* for 21 days. The vacuoles containing bacteria were labelled with lysosomal-associated membrane protein 1 (LAMP1; Fig. 1b, c),

indicating that these are endocytic compartments. In agreement with previous findings<sup>2</sup>, ~86% of bacteria within the liver sections were found in LAMP1<sup>+</sup> vacuoles (Fig. 1c). Approximately half of the *L. monocytogenes*-containing vacuoles were large (up to 7 µm in diameter), with only limited internal membranes (Fig. 1a, c). Therefore, we termed these compartments spacious *Listeria*-containing phagosomes (SLAPs). SLAPs often contained multiple intact bacteria, indicating that bacterial replication was occurring in these compartments.

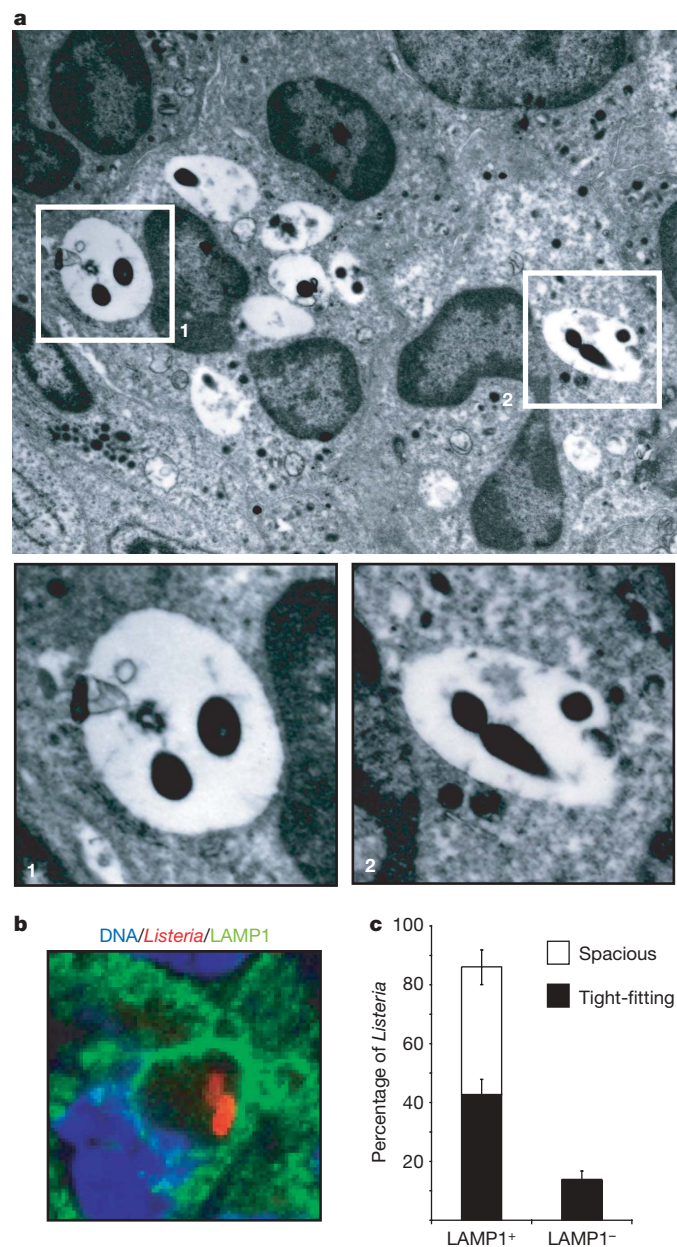
SLAP formation was also observed *in vitro* after *L. monocytogenes* infection of RAW 264.7 macrophages (Fig. 2a), J774 macrophages (data not shown) and primary bone-marrow derived macrophages (Supplementary Fig. 1). We used RAW 264.7 macrophages for the remainder of our *in vitro* studies. Although most *L. monocytogenes* escaped phagosomes and grew rapidly in the cytosol of RAW 264.7 macrophages as described previously<sup>1</sup>, we consistently observed a population of intracellular bacteria within vacuoles. The small percentage of intracellular bacteria that localized to SLAPs (~13% by 4 h post infection) was easily masked by robust replication of cytosolic bacteria and was difficult to observe without vacuolar markers. However, ~46% of infected cells formed SLAPs by this time in infection (Supplementary Fig. 2), and these structures were morphologically indistinguishable from the bacteria-containing compartments formed during persistent infection *in vivo*. Therefore, to gain further insight into the possible mechanisms governing persistent infection by *L. monocytogenes*, we further characterized the SLAP phenotype *in vitro*.

SLAPs often contained multiple intact bacteria (Fig. 2a) and colocalized with LAMP1 (Fig. 2b), similar to those observed in SCID mice. SLAPs also labelled with the autophagy marker LC3 (Fig. 2b, c), suggesting a role for autophagy in the formation of these compartments. Most SLAPs did not contain the lysosomal enzyme cathepsin D. In contrast, significant amounts of cathepsin D were observed in phagosomes containing bacteria killed with paraformaldehyde (PFA; Fig. 2d, e). These observations indicate that viable *L. monocytogenes* block SLAP maturation into degradative phagolysosomes.

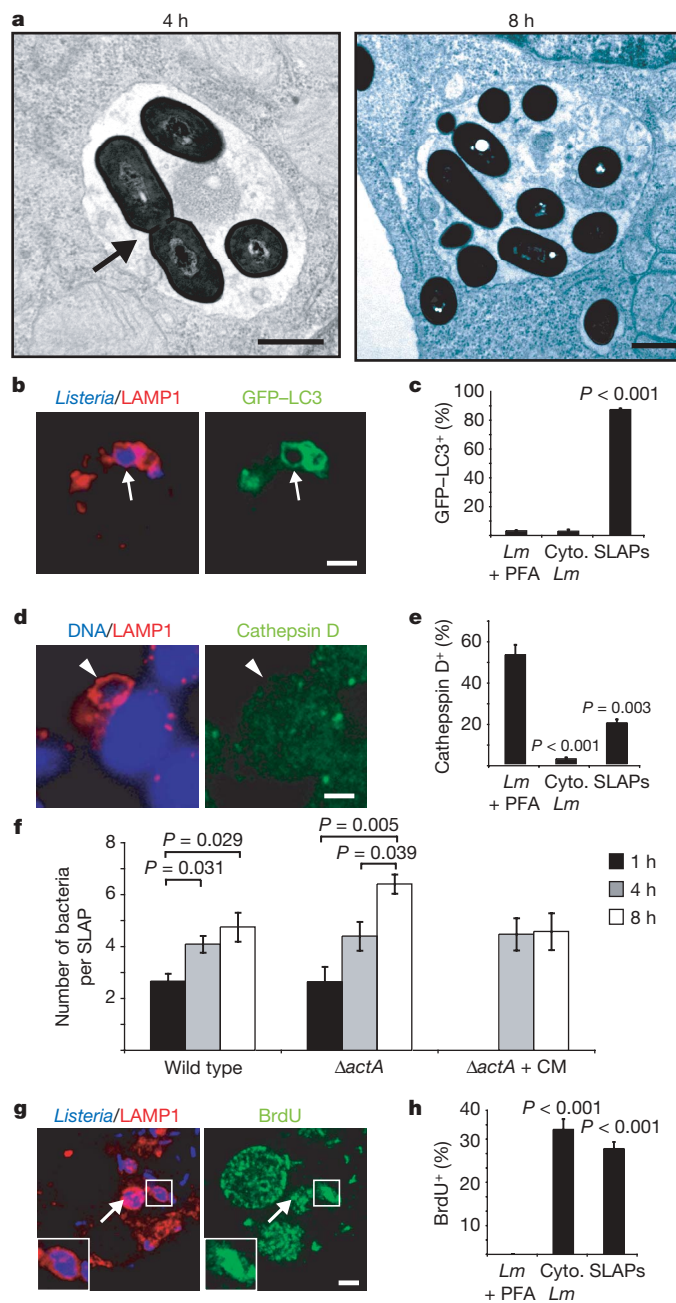
*L. monocytogenes* within SLAPs often exhibited septa (Fig. 2a, arrow), and the number of bacteria within these compartments increased over time (Fig. 2a, f). This increase in bacterial number within SLAPs was independent of cell-to-cell spread (because it also occurred with non-motile *actA* mutant bacteria) and required bacterial protein synthesis (Fig. 2f). These data suggest that bacteria replicate within SLAPs. To test this further, we stained *L. monocytogenes*-infected macrophages with bromodeoxyuridine (BrdU)—a thymidine analogue that is incorporated into replicating DNA. As shown in Fig. 2g and h, SLAPs often contained actively replicating bacteria that labelled with BrdU. It is possible that bacteria

<sup>1</sup>Cell Biology Program, Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada. <sup>2</sup>Department of Molecular Genetics, <sup>3</sup>Institute of Medical Science, University of Toronto, Toronto, Ontario M5S 1A8, Canada. <sup>4</sup>Department of Microbiology and Molecular Genetics, Harvard Medical School, Boston, Massachusetts 02115-6092, USA.

enter SLAPs after replication in the cytosol. However, non-motile *actA* mutant bacteria within SLAPs did not label with ubiquitin, which normally occurs when these bacteria are exposed to the cytosol<sup>8</sup> (Supplementary Fig. 3a). Also, monomeric red fluorescent protein expressed in the cytosol was not observed within SLAPs (Supplementary Fig. 3b), indicating that cytosolic contents are not delivered to these structures. Therefore, it seems that bacteria within SLAPs are not delivered from the cytosol, but may arise from a viable population that does not escape from the primary phagosome. Multiple bacteria-containing phagosomes may fuse together to form SLAPs. However, treatment of cells with either cytochalasin D or



**Figure 1** | *L. monocytogenes* colonize SLAPs during chronic infection of SCID mice. **a**, Mice were infected for 21 days and liver granulomas analysed by transmission electron microscopy (TEM). Shown are spacious vacuoles (SLAPs) containing multiple bacteria. Magnification,  $\times 5,200$ . Region 1 and 2 (white boxes) are enlarged in the lower panels. **b**, SCID mice were infected as in **a** and liver sections stained for LAMP1 (green), bacteria (red) and DNA (blue). Shown is a LAMP1<sup>+</sup> SLAP. **c**, The percentage of bacteria in LAMP1<sup>+</sup> compartments was quantified and characterized as either spacious or tight-fitting. Mean  $\pm$  s.e.m. for three mice examined. The image in **a** (from ref. 2) and the tissue sections were provided by E. Unanue.

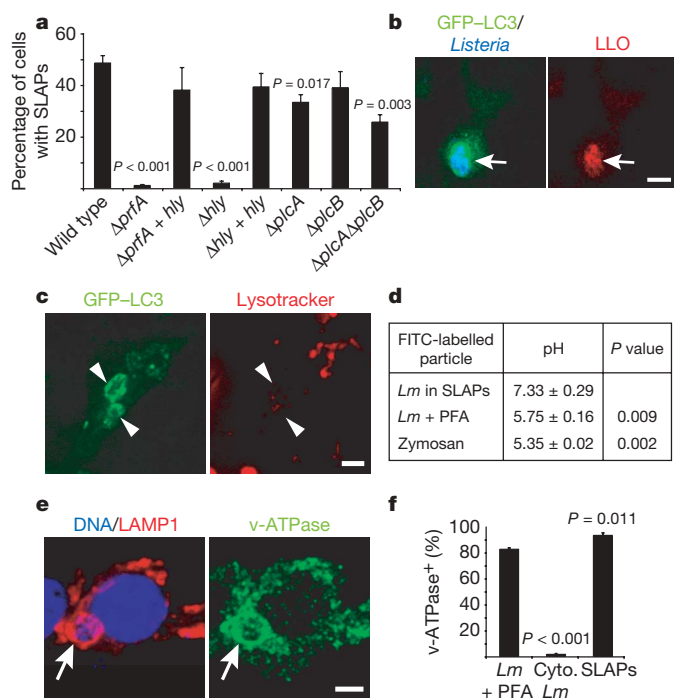


**Figure 2** | *L. monocytogenes* replicate slowly in SLAPs during *in vitro* infection of macrophages. **a**, RAW 264.7 macrophages infected for 4 or 8 h were analysed by TEM. Shown are spacious vacuoles (SLAPs) containing multiple bacteria. The arrow indicates septum of dividing bacteria. Scale bars, 0.5  $\mu$ m. **b**, The arrow indicates a LAMP1<sup>+</sup> SLAP colocalizing with green fluorescent protein (GFP)–LC3 in cells infected for 4 h. **d**, The arrowhead indicates a LAMP1<sup>+</sup> SLAP devoid of cathepsin D in cells infected for 4 h. Scale bars, 5  $\mu$ m. **c**, **e**, The percentage of GFP-LC3<sup>+</sup> (**c**) or cathepsin D<sup>+</sup> (**e**) SLAPs was quantified, and compared to GFP-LC3 or cathepsin D colocalization with cytosolic (Cyto.) bacteria (actin<sup>+</sup> or LAMP1<sup>+</sup>) or PFA-killed bacteria in phagosomes (LAMP1<sup>+</sup>). Mean  $\pm$  s.e.m. for three independent experiments. *P* values for conditions significantly different from PFA-killed bacteria are shown. **f**, GFP-LC3-transfected macrophages were infected with wild-type or  $\Delta actA$  bacteria. Where indicated, chloramphenicol (CM) was added to the media at 3 h post infection. The number of bacteria per SLAP was quantified. Brackets indicate significant differences, and corresponding *P* values are shown. **g**, Macrophages were infected with wild-type bacteria for 7 h, pulsed with BrdU for 1 h, and stained for LAMP1 (red), bacteria (blue) and BrdU (green). Magnified images and the arrow indicate SLAPs containing actively replicating bacteria (BrdU<sup>+</sup>). **h**, The percentage of BrdU<sup>+</sup> bacteria in SLAPs, compared to cytosolic and PFA-killed bacteria, was quantified as in **c**.

nocodazole—inhibitors that disrupt the actin and microtubule cytoskeletons, respectively, and thus impair membrane traffic—did not affect the number of bacteria within SLAPs (Supplementary Fig. 3c). Therefore, our data are consistent with bacterial replication within SLAPs.

SLAP formation required continuous bacterial protein synthesis (Supplementary Fig. 2). Therefore, we tested for bacterial virulence factors involved in the formation of these structures. PrfA is a main transcriptional regulator of virulence genes in *L. monocytogenes*<sup>9</sup>. A *prfA* mutant did not form SLAPs (Fig. 3a). An *hly*-deletion mutant, which does not express LLO, also did not form SLAPs, indicating that LLO is necessary for the formation of these compartments (Fig. 3a). Two bacterial phospholipase Cs (PLCs) encoded by *plcA* and *plcB* assist LLO in mediating bacterial escape from the phagosome<sup>1</sup>. However, bacterial mutants of these genes had only minor defects in SLAP formation. *hly* expression in a *prfA* mutant ( $\Delta prfA + hly$ ) rescued SLAP formation, indicating that LLO is sufficient for the formation of SLAPs (Fig. 3a). LLO was expressed within SLAPs, as shown by specific staining with monoclonal antibodies (Fig. 3b). Therefore, a localized effect of LLO on the vacuole seems to allow bacterial replication within SLAPs.

LLO is known to uncouple pH gradients of the primary phagosome by creating small pores in the phagosomal membrane<sup>10</sup>. This is thought to allow a window of opportunity for LLO- and PLC-mediated lysis of the phagosome, as well as bacterial escape into the cytosol<sup>10</sup>. Because LLO was both sufficient and necessary for SLAP formation and was acting within SLAPs, we hypothesized that



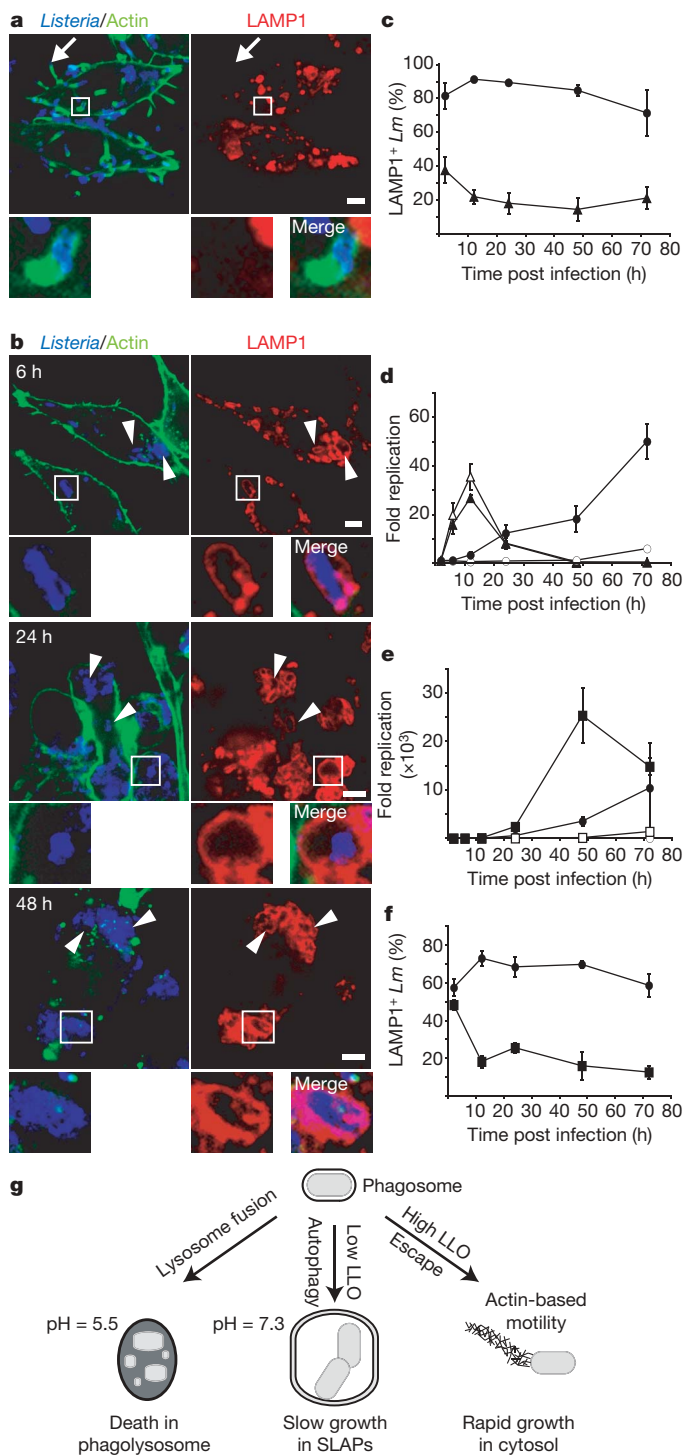
**Figure 3 | SLAP formation requires bacterial LLO expression.** **a**, GFP-LC3-transfected macrophages were infected for 4 h, and the percentage of infected cells exhibiting SLAPs was quantified. Mean  $\pm$  s.e.m. for three independent experiments. *P* values for strains with significant differences from wild-type levels are shown. **b**, The arrow indicates a GFP-LC3<sup>+</sup> SLAP with internal LLO expression in cells infected for 4 h. Scale bar, 5  $\mu$ m. **c**, Arrowheads indicate GFP-LC3<sup>+</sup> SLAPs devoid of Lysotracker Red in cells infected for 4 h. **d**, The pH of FITC-labelled bacteria in SLAPs, PFA-killed bacteria or zymosan particles was determined by ratiometric imaging. *P* values compared to bacteria in SLAPs are shown. **e**, The arrow indicates a LAMP1<sup>+</sup> SLAP colocalizing with v-ATPase staining in cells infected for 4 h. Scale bar, 5  $\mu$ m. **f**, The percentage of v-ATPase<sup>+</sup> SLAPs was quantified as in Fig. 2c. Mean  $\pm$  s.e.m. for three independent experiments. *P* values for conditions significantly different from PFA-killed bacteria are shown.

LLO might also uncouple pH gradients across SLAP membranes. Consistent with this hypothesis, most ( $84 \pm 4.8\%$ ) SLAPs were negative for the acidotropic dye LysoTracker Red (Fig. 3c). To measure the pH of SLAPs directly, we used ratiometric imaging of bacteria pre-labelled with the pH-sensitive dye fluorescein isothiocyanate (FITC)<sup>11</sup>. As shown in Fig. 3d and Supplementary Fig. 4, SLAPs were found to be neutral compartments (average pH  $7.3 \pm 0.29$ ). Phagosomes containing PFA-killed bacteria or zymosan particles acidified to an average pH of  $5.8 \pm 0.16$  and  $5.4 \pm 0.02$ , respectively (Fig. 3d and Supplementary Fig. 4b), consistent with previous studies of phagolysosomes<sup>12</sup>. However, SLAPs were positive for v-ATPase staining (Fig. 3e, f), indicating that the proton pump was present on these compartments. These results are consistent with LLO forming small pores in the SLAP membrane to uncouple the pH gradient. Acidification is known to be required for phagosome and autophagosome maturation<sup>13,14</sup>. Therefore, by blocking acidification of SLAPs, LLO effectively blocks fusion of this compartment with lysosomes, allowing a population of bacteria to replicate within vacuoles.

LLO expression is required for SLAP formation. However, *L. monocytogenes* within SLAPs seem to arise from a bacterial population that does not successfully escape from the primary phagosome. Therefore, bacteria within SLAPs may have reduced LLO expression or inefficient LLO activity. It has been shown previously that LLO activity is impaired by innate immune factors in activated macrophages, and is inefficient in LAMP1<sup>+</sup> compartments and alkaline environments<sup>5,6,10,15</sup>. Therefore, we hypothesized that experimentally reducing LLO expression would block *L. monocytogenes* entry into the cytosol but promote bacterial replication within SLAPs. To test this, we used an LLO-deficient (*hly* mutant) of *L. monocytogenes* that expresses LLO under a tightly controlled isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG)-inducible promoter (iLLO)<sup>16</sup>. With maximal induction, the haemolytic activity of the iLLO strain is approximately 33% that of wild-type *L. monocytogenes*<sup>16</sup>.

We compared the intracellular replication of the iLLO strain in macrophages to that of wild-type bacteria. As expected, wild-type *L. monocytogenes* exhibited rapid replication and most bacteria were LAMP1<sup>−</sup> (Fig. 4a, c). Consistent with localization in the cytosol, we observed wild-type bacteria associated with actin 'comet tails' and undergoing actin-based motility (Fig. 4a). Intracellular numbers of wild-type bacteria peaked at 12 h post infection and then declined (Fig. 4d). In contrast, the iLLO strain grew slowly in macrophages, approaching the same intracellular numbers as wild-type *L. monocytogenes* only after 48 to 72 h post infection (Fig. 4b, d). Replication of iLLO bacteria required continuous induction of LLO expression (Fig. 4d), and removal of IPTG at 12 h post infection blocked subsequent growth (data not shown). Most iLLO bacteria remained LAMP1<sup>+</sup> (Fig. 4b, c) and did not display evidence of having entered the cytosol throughout the course of infection (Supplementary Fig. 5). Therefore, LLO permits replication of *L. monocytogenes* within vacuoles when its activity is not sufficient to drive escape into the cytosol.

SLAPs were positive for the autophagy marker LC3 (Fig. 2b, c), and we have shown previously that *L. monocytogenes* can be targeted by autophagy early in infection<sup>17</sup>. Therefore, we hypothesized that autophagy may be involved in SLAP formation. In support of this, we found that autophagy inhibitors blocked SLAP formation (Supplementary Fig. 6). Because SLAP formation required LLO (Fig. 3a), we hypothesized that autophagy targets damaged phagosomes to prevent bacterial escape into the cytosol. To test this hypothesis, we infected autophagy-deficient (*Atg5*<sup>−/−</sup>) mouse embryonic fibroblasts (MEFs)<sup>18</sup> with iLLO bacteria. On induction of LLO expression, these bacteria grew rapidly in *Atg5*<sup>−/−</sup> MEFs (Fig. 4e). Under these conditions, most bacteria did not colocalize with LAMP1 (Fig. 4f). In contrast, autophagy-competent MEFs maintained iLLO bacteria within LAMP1<sup>+</sup> vacuoles and delayed the kinetics of their replication (Fig. 4e, f). In the absence of induction, iLLO bacteria did not replicate in either cell type (Fig. 4e). In control



experiments, ~90% of wild-type bacteria were LAMP1<sup>+</sup> throughout infection in both autophagy-competent and autophagy-deficient MEFs (data not shown), demonstrating that normal LLO expression is sufficient to drive phagosomal escape in this cell type. These studies demonstrate that autophagy restricts *L. monocytogenes* replication to LAMP1<sup>+</sup> vacuoles under conditions when LLO expression is impaired.

Here we present the first study of mechanisms governing *L. monocytogenes* replication in vacuoles of host cells. We characterize a novel compartment, the SLAP, which is permissive for bacterial replication. *L. monocytogenes* replicate rapidly in the cytosol (doubling time of approximately 40 min<sup>1,19</sup>) but slowly within SLAPs (doubling time of approximately 8 h, Fig. 2f). It is not known whether the

**Figure 4 | Impaired LLO expression allows slow bacterial replication within vacuoles.** **a**, Arrows indicate LAMP1<sup>+</sup> wild-type bacteria with actin 'comet tails' in cells infected for 6 h. The boxed region is magnified in the bottom left panels. The merged image for this region is shown at the bottom right. Scale bar, 5  $\mu$ m. **b**, Macrophages were infected with IPTG-induced iLLO bacteria. Arrowheads indicate actin<sup>+</sup> iLLO bacteria within LAMP1<sup>+</sup> vacuoles. **c**, Macrophages were infected with wild-type (triangles) or IPTG-induced iLLO (circles) bacteria, and the percentage of LAMP1<sup>+</sup> bacteria was quantified. Mean  $\pm$  s.e.m. for three independent experiments. **d**, Macrophages were infected as in **c** with or without IPTG induction. Intracellular bacterial replication was determined using a gentamicin-protection assay. Shown is fold replication compared to 2 h post infection. Clear triangles, wild type; filled triangles, wild type + IPTG; clear circles, iLLO; filled circles, iLLO + IPTG. Mean  $\pm$  s.e.m. or range for three (wild type, wild type + IPTG, iLLO + IPTG) or two (iLLO–IPTG) independent experiments, respectively. **e**, Wild-type or *Atg5*<sup>−/−</sup> MEFs were infected with iLLO bacteria, and intracellular bacterial replication was determined as in **d**. Mean  $\pm$  s.e.m. for three independent experiments. Clear circles, wild-type MEFs; filled circles, wild-type MEFs + IPTG; clear squares, *Atg5*<sup>−/−</sup> MEFs; filled squares, *Atg5*<sup>−/−</sup> MEFs + IPTG. **f**, Wild-type (circles) or *Atg5*<sup>−/−</sup> (squares) MEFs were infected as in **e** with IPTG induction. The percentage of LAMP1<sup>+</sup> bacteria was quantified as in **c**. Mean  $\pm$  s.e.m. for three independent experiments. **g**, Model of the different fates of *L. monocytogenes* in host cells. High LLO activity allows bacterial escape from phagosomes. Under conditions where LLO activity is not sufficient to drive escape (low LLO), autophagy maintains bacteria within non-degradative vacuoles (SLAPs) that allow slow bacterial growth. Bacteria can also be degraded in phagolysosomes.

mechanisms governing SLAP formation *in vitro* are the same as those involved in the morphogenesis of bacteria-containing vacuoles during infection of SCID mice<sup>2</sup>. However, the fact that these structures both label with endocytic markers, are morphologically comparable and contain multiple bacteria suggests that the mechanisms of formation are similar.

Bacterial replication within SLAPs seems to represent a delicate balance between virulence factors of the pathogen and innate immune mechanisms of the infected cell. LLO was necessary and sufficient for *L. monocytogenes* replication within SLAPs. Therefore, LLO can be ascribed several key virulence functions: blocking nascent phagosome maturation by uncoupling the pH gradient across the phagosomal membrane<sup>4</sup>; mediating phagosome escape<sup>1</sup>; and triggering autophagy of damaged phagosomes and blocking their maturation, leading to SLAP formation and bacterial growth in vacuoles (this study). Therefore, differential LLO activities seem to give rise to different fates of *L. monocytogenes* within host cells (Fig. 4g). Our studies also demonstrate that a host cellular process, namely autophagy, maintains *L. monocytogenes* in vacuoles, particularly when LLO activity is impaired. SLAPs seem to represent a 'stalemate' for *L. monocytogenes* infection. The host cell is able to sustain viability by preventing bacterial colonization of the cytosol, but is unable to eradicate the pathogen. At the same time, the pathogen is able to replicate in SLAPs but at a reduced rate compared to that in its favoured niche, the cytosol. It remains to be seen whether other bacterial pathogens that express cholesterol-dependent cytolysins<sup>3</sup> utilize these toxins in a manner similar to LLO to promote their growth in vacuoles in host cells.

## METHODS SUMMARY

The *L. monocytogenes* strains used are listed in Methods. Infections of C.B-17/ICR SCID mice were performed as described previously<sup>2</sup>. *In vitro* infections were performed in the presence of gentamicin to prevent extracellular growth at a multiplicity of infection (MOI) of 10 for RAW 264.7 macrophages and an MOI of 50 for MEFs (unless otherwise stated in Methods). LLO expression in iLLO bacteria was induced as described previously<sup>16</sup>. Any pharmacological agents used are listed in Methods.

TEM and immunofluorescence were performed as described<sup>8,20,21</sup>. Antibodies and dyes used are listed in Methods. Antigen retrieval (boiling in 10 mM sodium citrate) was performed for tissue and BrdU staining.

Most colocalization quantifications were performed by direct visualization on a Leica DMIRE2 epifluorescence microscope. All images shown are confocal *z* slices taken using a Zeiss Axiovert confocal microscope and LSM 510 software. Live imaging was performed on a Leica DMIRE2 inverted confocal microscope with a Hamamatsu Back-Thinned EM-CCD camera and spinning disk scan head. Volocity software (Improvision) was used to analyse images and to assemble *z* slices. Figure assembly was done using Adobe PhotoShop and Adobe Illustrator.

For pH measurements, wild-type bacteria, PFA-killed IgG-opsonized bacteria or zymosan particles were covalently labelled with 0.5 mg ml<sup>-1</sup> FITC and added to RFP-LC3-transfected RAW 264.7 cells for 45 min or 4 h as indicated. Ratiometric imaging was performed as described previously<sup>11</sup> on a Leica DM IRB microscope with 485 nm and 438 nm excitation filters and a Cascade II CCD camera. Where appropriate, a corresponding red channel image (545 nm excitation) was acquired. Calibrations were performed with isotonic K<sup>+</sup> solutions of known pH values containing 1 µM nigericin.

The mean ± standard error (s.e.m.) is shown in figures, and *P* values were calculated using a two-tailed two-sample equal variance Student's *t*-test. A *P* value of less than 0.05 was determined to be statistically significant.

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Received 25 September; accepted 13 November 2007.**

- Portnoy, D. A., Auerbuch, V. & Glomski, I. J. The cell biology of *Listeria monocytogenes* infection: the intersection of bacterial pathogenesis and cell-mediated immunity. *J. Cell Biol.* **158**, 409–414 (2002).
- Bhardwaj, V., Kanagawa, O., Swanson, P. E. & Unanue, E. R. Chronic *Listeria* infection in SCID mice: requirements for the carrier state and the dual role of T cells in transferring protection or suppression. *J. Immunol.* **160**, 376–384 (1998).
- Kayal, S. & Charbit, A. Listeriolysin O: a key protein of *Listeria monocytogenes* with multiple functions. *FEMS Microbiol. Rev.* **30**, 514–529 (2006).
- Shaughnessy, L. M., Hoppe, A. D., Christensen, K. A. & Swanson, J. A. Membrane perforations inhibit lysosome fusion by altering pH and calcium in *Listeria monocytogenes* vacuoles. *Cell. Microbiol.* **8**, 781–792 (2006).
- Myers, J. T., Tsang, A. W. & Swanson, J. A. Localized reactive oxygen and nitrogen intermediates inhibit escape of *Listeria monocytogenes* from vacuoles in activated macrophages. *J. Immunol.* **171**, 5447–5453 (2003).
- del Cerro-Vadillo, E. *et al.* Cutting edge: a novel nonoxidative phagosomal mechanism exerted by cathepsin-D controls *Listeria monocytogenes* intracellular growth. *J. Immunol.* **176**, 1321–1325 (2006).
- Pamer, E. G. Immune responses to *Listeria monocytogenes*. *Nature Rev. Immunol.* **4**, 812–823 (2004).
- Perrin, A. J., Jiang, X., Birmingham, C. L., So, N. S. & Brumell, J. H. Recognition of bacteria in the cytosol of Mammalian cells by the ubiquitin system. *Curr. Biol.* **14**, 806–811 (2004).
- Hamon, M., Bierre, H. & Cossart, P. *Listeria monocytogenes*: a multifaceted model. *Nature Rev. Microbiol.* **4**, 423–434 (2006).
- Henry, R. *et al.* Cytolysin-dependent delay of vacuole maturation in macrophages infected with *Listeria monocytogenes*. *Cell. Microbiol.* **8**, 107–119 (2006).
- Jankowski, A., Scott, C. C. & Grinstein, S. Determinants of the phagosomal pH in neutrophils. *J. Biol. Chem.* **277**, 6059–6066 (2002).
- Hackam, D. J. *et al.* Regulation of phagosomal acidification. Differential targeting of Na<sup>+</sup>/H<sup>+</sup> exchangers, Na<sup>+</sup>/K<sup>+</sup>-ATPases, and vacuolar-type H<sup>+</sup>-ATPases. *J. Biol. Chem.* **272**, 29810–29820 (1997).
- Gordon, A. H., Hart, P. D. & Young, M. R. Ammonia inhibits phagosome-lysosome fusion in macrophages. *Nature* **286**, 79–80 (1980).
- Yamamoto, A. *et al.* Bafilomycin A1 prevents maturation of autophagic vacuoles by inhibiting fusion between autophagosomes and lysosomes in rat hepatoma cell line, H-4-II-E cells. *Cell Struct. Funct.* **23**, 33–42 (1998).
- Beauregard, K. E., Lee, K. D., Collier, R. J. & Swanson, J. A. pH-dependent perforation of macrophage phagosomes by listeriolysin O from *Listeria monocytogenes*. *J. Exp. Med.* **186**, 1159–1163 (1997).
- Alberti-Segui, C., Goeden, K. R. & Higgins, D. E. Differential function of *Listeria monocytogenes* listeriolysin O and phospholipases C in vacuolar dissolution following cell-to-cell spread. *Cell. Microbiol.* **9**, 179–195 (2007).
- Birmingham, C. L. *et al.* *Listeria monocytogenes* evades killing by autophagy during colonization of host cells. *Autophagy* **3**, 442–451 (2007).
- Kuma, A. *et al.* The role of autophagy during the early neonatal starvation period. *Nature* **432**, 1032–1036 (2004).
- de Chastellier, C. & Berche, P. Fate of *Listeria monocytogenes* in murine macrophages: evidence for simultaneous killing and survival of intracellular bacteria. *Infect. Immun.* **62**, 543–553 (1994).
- Brumell, J. H., Rosenberger, C. M., Gotto, G. T., Marcus, S. L. & Finlay, B. B. SifA permits survival and replication of *Salmonella typhimurium* in murine macrophages. *Cell. Microbiol.* **3**, 75–84 (2001).
- Kaniuk, N. A. *et al.* Ubiquitinated-protein aggregates form in pancreatic beta-cells during diabetes-induced oxidative stress and are regulated by autophagy. *Diabetes* **56**, 930–939 (2007).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** J.H.B. holds an Investigators in Pathogenesis of Infectious Disease Award from the Burroughs Wellcome Fund and is the recipient of the Premier's Research Excellence Award from the Ontario Ministry of Economic Development and Trade and the Boehringer Ingelheim (Canada) Young Investigator Award in Biological Sciences. Laboratory infrastructure was provided by a New Opportunities Fund from the Canadian Foundation for Innovation and the Ontario Innovation Trust. C.L.B. holds a Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada. N.A.K. holds a CAG/CIHR/Axcan Pharma fellowship from the Canadian Association of Gastroenterology. B.E.S. is supported by Canadian Institutes of Health Research and McLaughlin Centre for Molecular Medicine MD/PhD studentships. We are grateful to E. R. Unanue for performing *in vivo* infections of mice and providing tissue sections and electron micrographs. We thank D. Brown, P. Cossart, J. Danska, E. Gouin, S. Grinstein, N. Jones, N. Mizushima, D. Portnoy, and T. Yoshimori for providing reagents and suggestions. We also thank M. Woodside, P. Paroutis and R. Temkin for assistance with microscopy.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence and requests for materials should be addressed to J.H.B. ([john.brumell@sickkids.ca](mailto:john.brumell@sickkids.ca)).

## METHODS

**In vivo infections.** Infections of C.B-17/ICR SCID mice were performed as described previously<sup>2</sup>.

**Cell culture and bacterial strains.** RAW 264.7 macrophages and wild-type and *Atg5*<sup>-/-</sup> MEFs<sup>18</sup> were maintained in DMEM medium (HyClone) with 10% FBS (Wisent) at 37 °C in 5% CO<sub>2</sub> without antibiotics. Bone-marrow derived macrophages harvested from NOD mice were provided by J. Danska and maintained in supplemented growth media containing 10 ng ml<sup>-1</sup> granulocyte monocyte colony stimulating factor for 6–7 days before use.

*L. monocytogenes* were grown in brain-heart infusion (BHI) broth and the following strains used: wild-type 10403S (ref. 22),  $\Delta actA$  (DP-L3078, ref. 23),  $\Delta prfA$  (DP-L4137, ref. 24),  $\Delta prfA + hly$  (DH-L919, ref. 17),  $\Delta hly$  (DP-L2161, ref. 25),  $\Delta hly + hly$  (DP-L4818, ref. 26),  $\Delta plcA$  (DP-L1552, ref. 27),  $\Delta plcB$  (DP-L1935, ref. 28),  $\Delta plcA\Delta plcB$  (DP-L1936, ref. 28), iLLO (DH-L1239, ref. 16) and  $\Delta actA$  iLLO (DH-L1257, ref. 16).

**In vitro infections.** Most infections of macrophages were performed as described previously<sup>17</sup>. An MOI of 10 was used, except for Fig. 4b in which an MOI of 100 was used. Infection of macrophages with iLLO bacteria was performed as described previously<sup>16</sup>. Bacteria were induced with 0.5 mM IPTG for 2 h before infection, and 10 mM IPTG was maintained in the media for the duration of the experiment. For infection of MEFs with iLLO *L. monocytogenes*, bacteria were grown overnight at room temperature (~22 °C). Cells were infected as above at an MOI of 50, and gentamicin added to the media at 1 h post infection. Gentamicin-protected intracellular replication assays were performed as described previously<sup>17</sup>. Fold replication was determined by dividing CFUs at the desired time by CFUs at 2 h post infection. For IPTG pulse-chase experiments, macrophages were infected with IPTG-induced iLLO bacteria as above. At 12 h, cells were extensively washed with PBS, and media without IPTG was added for the remainder of the experiment.

To kill bacteria with PFA, bacteria grown overnight in BHI broth were harvested, washed in PBS and rotated at room temperature for 30 min in 13% PFA. Bacterial killing was confirmed by plating on growth plates.

Chloramphenicol (200  $\mu$ g ml<sup>-1</sup>) was added at 3 h post infection. Autophagy inhibitors wortmannin (Sigma; 100 nM), 3-methyladenine (Sigma; 10 mM) and LY294002 (Sigma; 100  $\mu$ M) were added at 30 min post infection. Nocodazole (Sigma; 5  $\mu$ M) and cytochalasin D (Sigma; 10  $\mu$ M) were added at 1 h post infection.

**Transmission electron microscopy, immunofluorescence and transfection.** For TEM, cells were fixed in 2% glutaraldehyde overnight (~16 h) at room temperature and processed as described previously<sup>8</sup>.

Immunofluorescence of tissue sections was performed as described previously<sup>21</sup> with an antigen-retrieval step (boiling in 10 mM sodium citrate buffer (pH 6.0) for 30 min). For immunofluorescence of tissue culture cells, cells were fixed with 2.5% PFA for 10 min at 37 °C, except for LLO staining (methanol at -20 °C for 10 min) and cathepsin D and v-ATPase staining (post-fix with methanol at -20 °C for 10 min). Permeabilization and blocking were performed with 0.2% saponin and 10% normal goat serum overnight at 4 °C. Staining was performed as described previously<sup>20</sup>. Quantifications were performed on a Leica DMIRE2 epifluorescence microscope. All images shown are confocal z slices from a Zeiss Axiovert confocal microscope using LSM 510 software.

The following antibodies and dyes were used: rabbit anti-*L. monocytogenes* (generated as described previously<sup>29</sup>), rat anti-LAMP1 (Developmental Studies Hybridoma Bank under the auspices of the NICHD and maintained by the University of Iowa), mouse anti-ubiquitinated proteins (Affiniti Research Products Ltd), mouse anti-LLO (generated as described previously<sup>30</sup>), rabbit anti-cathepsin D (Scripps Research Institute), rabbit anti-v-ATPase (from D. Brown) and phalloidin conjugated to AlexaFluor 488 or 568 (Molecular Probes). All secondary antibodies used were AlexaFluor conjugates (Molecular Probes). DAPI (Molecular Probes) was used according to the manufacturer's instructions.

BrdU was added to the media for 1 h and cells were fixed in methanol at -20 °C for 20 min. Antigen retrieval was performed (boiling in 10 mM sodium citrate buffer (pH 6.0) for 10 min) before samples were permeabilized/blocked in 5% BSA with 0.2% saponin. Staining was performed as above. Goat anti-BrdU was from J. Gordon.

Cells were transfected with FuGene 6 (Roche Diagnostics) or ExGen 500 (Fermentas) according to the manufacturers' instructions. GFP-LC3 and the plasmid expressing monomeric red fluorescent protein were generated as described previously<sup>31,32</sup>.

**Lysotracker labelling.** Coverslips seeded with cells were maintained in imaging chambers in RPMI media with HEPES (without bicarbonate) (HyClone). Bacteria grown overnight at 37 °C shaking were diluted 1:10, subcultured for 2 h, and added to the cells at an MOI of 30. Cells were incubated at 37 °C with 5%

CO<sub>2</sub>. At 30 min, extracellular bacteria were removed by washing and gentamicin was added to the media. At 4 h, live imaging was performed on a Leica DMIRE2 inverted confocal microscope with a Hamamatsu Back-Thinned EM-CCD camera and spinning disk scan head. Volocity software (Improvision) was used. LysoTracker Red (Molecular Probes; 100 nM) was loaded into cells at the time of infection.

**pH measurements.** Bacteria were labelled with 0.5 mg ml<sup>-1</sup> FITC in PBS (pH 8.9) for 20 min shaking at 37 °C, followed by extensive washing. For PFA-killed samples, labelled *L. monocytogenes* were treated with PFA as above and were opsonized in 10 mg ml<sup>-1</sup> human IgG by rotating for 1 h at room temperature. Zymosan (Molecular Probes) particles were incubated with 0.5 mg ml<sup>-1</sup> FITC and were IgG-opsonized using zymosan opsonizing reagent (Molecular Probes).

RFP-LC3-transfected RAW cells were used for all experiments. Live bacterial invasion was performed with FITC-*L. monocytogenes* as above. FITC-PFA-killed bacteria were centrifuged onto cells at 250g for 5 min at 4 °C. Cells were incubated at 37 °C to allow phagocytosis. At 30 min, cells were placed on ice, and goat anti-human AlexaFluor 568 (Molecular Probes) was added to the coverslip for 2.5 min to label extracellular bacteria. The antibody was washed off and cells incubated at 37 °C for a further 15 min (total of 45 min after addition of PFA-killed bacteria). FITC-zymosan were centrifuged onto cells at 550g for 1 min. Cells were incubated for 5 min at 37 °C to allow for particle internalization, and then vigorously washed to remove uninternalized particles. Phagosome maturation was allowed to proceed for 45 min.

Samples were maintained at 37 °C and imaged with a Leica DM IRB microscope. pH was measured by fluorescence ratiometric imaging. Light was transmitted alternately through 485  $\pm$  10 nm and 438  $\pm$  12 nm excitation filters and directed with a 505 nm dichroic mirror. Emitted light filtered with a 535  $\pm$  20 nm emission filter was captured by a Cascade II CCD camera. The filter wheel and camera were controlled with Metafluor software (Molecular Devices). Where appropriate, a corresponding red channel image (545  $\pm$  15 nm excitation filter, 570 nm dichroic mirror, and 610  $\pm$  37 nm emission filter) was acquired to discern either extracellular PFA-killed bacteria or RFP-LC3 signal.

*In situ* calibrations were performed by sequentially bathing the cells in isotonic K<sup>+</sup> solutions (145 mM KCl, 10 mM glucose, 1 mM MgCl<sub>2</sub>, 1 mM CaCl<sub>2</sub> and 20 mM of either HEPES or MES or acetate) buffered to pH values from 5.0 to 7.5 and containing 1  $\mu$ M nigericin. The resulting fluorescence intensity ratio (490/440 nm) as a function of pH was fit to a Boltzmann sigmoid and was used to interpolate pH values from the experimental ratio data.

All image analysis was carried out using background-subtracted fluorescence intensities for user-defined regions-of-interest with MetaFluor software. PFA-killed-bacteria-containing phagosomes were identified by the lack of extracellular secondary antibody (red) staining. SLAPs were identified as large RFP-LC3<sup>+</sup> structures that colocalized with bacteria.

**Statistics.** Colocalization quantifications were performed by direct visualization on a Leica DMIRE2 epifluorescence microscope (except confocal z slices were used for Fig. 4c). At least 100 bacteria, cells or SLAPs were counted for each condition in each experiment. For pH measurements, 15–25 SLAPs, at least 60 PFA-killed bacteria or at least 170 zymosan particles were analysed. At least three independent experiments were performed unless otherwise indicated (two independent experiments were performed for LysoTracker studies). The mean  $\pm$  s.e.m. is shown in figures unless otherwise indicated, and *P* values were calculated using a two-tailed two-sample equal variance Student's *t*-test. A *P* value of less than 0.05 was determined to be statistically significant.

22. Bishop, D. K. & Hinrichs, D. J. Adoptive transfer of immunity to *Listeria monocytogenes*. The influence of *in vitro* stimulation on lymphocyte subset requirements. *J. Immunol.* **139**, 2005–2009 (1987).
23. Skoble, J., Portnoy, D. A. & Welch, M. D. Three regions within ActA promote Arp2/3 complex-mediated actin nucleation and *Listeria monocytogenes* motility. *J. Cell Biol.* **150**, 527–538 (2000).
24. Cheng, L. W. & Portnoy, D. A. *Drosophila* S2 cells: an alternative infection model for *Listeria monocytogenes*. *Cell. Microbiol.* **5**, 875–885 (2003).
25. Jones, S. & Portnoy, D. A. Characterization of *Listeria monocytogenes* pathogenesis in a strain expressing perfringolysin O in place of listeriolysin O. *Infect. Immun.* **62**, 5608–5613 (1994).
26. Lauer, P., Chow, M. Y., Loessner, M. J., Portnoy, D. A. & Calendar, R. Construction, characterization, and use of two *Listeria monocytogenes* site-specific phage integration vectors. *J. Bacteriol.* **184**, 4177–4186 (2002).
27. Camilli, A., Tilney, L. G. & Portnoy, D. A. Dual roles of *plcA* in *Listeria monocytogenes* pathogenesis. *Mol. Microbiol.* **8**, 143–157 (1993).
28. Smith, G. A. *et al.* The two distinct phospholipases C of *Listeria monocytogenes* have overlapping roles in escape from a vacuole and cell-to-cell spread. *Infect. Immun.* **63**, 4231–4237 (1995).

29. Dramsi, S., Levi, S., Triller, A. & Cossart, P. Entry of *Listeria monocytogenes* into neurons occurs by cell-to-cell spread: an *in vitro* study. *Infect. Immun.* **66**, 4461–4468 (1998).
30. Nato, F. *et al.* Production and characterization of neutralizing and nonneutralizing monoclonal antibodies against listeriolysin O. *Infect. Immun.* **59**, 4641–4646 (1991).
31. Kabeya, Y. *et al.* LC3, a mammalian homologue of yeast Apg8p, is localized in autophagosome membranes after processing. *EMBO J.* **19**, 5720–5728 (2000).
32. Campbell, R. E. *et al.* A monomeric red fluorescent protein. *Proc. Natl Acad. Sci. USA* **99**, 7877–7882 (2002).

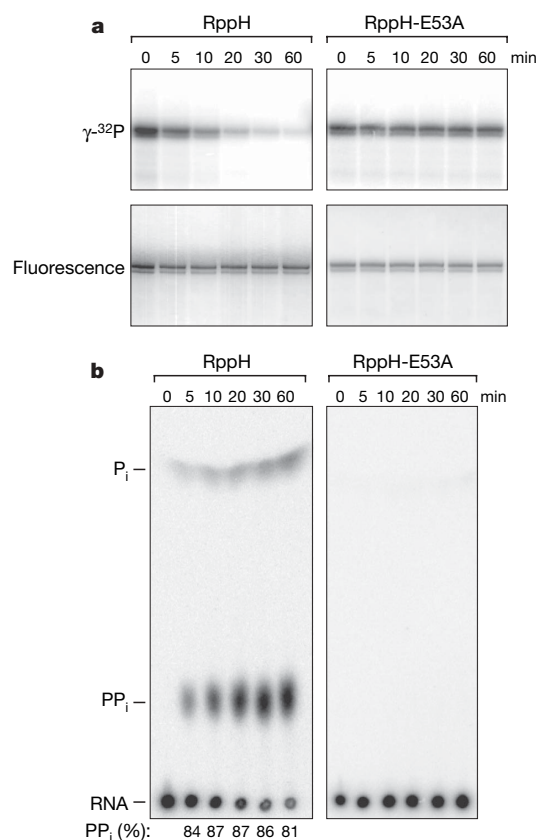
# The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal

Atilio Deana<sup>1</sup>, Helena Celesnik<sup>1</sup> & Joel G. Belasco<sup>1</sup>

The long-standing assumption that messenger RNA (mRNA) degradation in *Escherichia coli* begins with endonucleolytic cleavage has been challenged by the recent discovery that RNA decay can be triggered by a prior non-nucleolytic event that marks transcripts for rapid turnover: the rate-determining conversion of the 5' terminus from a triphosphate to a monophosphate<sup>1</sup>. This modification creates better substrates for the endonuclease RNase E, whose cleavage activity at internal sites is greatly enhanced when the RNA 5' end is monophosphorylated<sup>2,3</sup>. Moreover, it suggests an explanation for the influence of 5' termini on the endonucleolytic cleavage of primary transcripts, which are triphosphorylated<sup>4–8</sup>. However, no enzyme capable of removing pyrophosphate from RNA 5' ends has been identified in any bacterial species. Here we show that the *E. coli* protein RppH (formerly NudH/YgdP) is the RNA pyrophosphohydrolase that initiates mRNA decay by this 5'-end-dependent pathway. *In vitro*, RppH efficiently removes pyrophosphate from the 5' end of triphosphorylated RNA, irrespective of the identity of the 5'-terminal nucleotide. *In vivo*, it accelerates the degradation of hundreds of *E. coli* transcripts by converting their triphosphorylated 5' ends to a more labile monophosphorylated state that can stimulate subsequent ribonuclease cleavage. That the action of the pyrophosphohydrolase is impeded when the 5' end is structurally sequestered by a stem-loop helps to explain the stabilizing influence of 5'-terminal base pairing on mRNA lifetimes. Together, these findings suggest a possible basis for the effect of RppH and its orthologues on the invasiveness of bacterial pathogens. Interestingly, this master regulator of 5'-end-dependent mRNA degradation in *E. coli* not only catalyses a process functionally reminiscent of eukaryotic mRNA decapping but also bears an evolutionary relationship to the eukaryotic decapping enzyme Dcp2.

We reasoned that a protein with RNA pyrophosphohydrolase activity might previously have been identified as an enzyme able to remove pyrophosphate from mononucleotides. Because several members of the Nudix protein family have been shown to possess mononucleotide pyrophosphohydrolase activity *in vitro*<sup>9</sup>, we purified 12 Nudix proteins from *E. coli* and tested them individually for their ability to remove pyrophosphate from the 5' end of triphosphorylated RNA. This screening revealed that RppH functions *in vitro* as an efficient RNA pyrophosphohydrolase. When added to RNA bearing a 5'-terminal  $\gamma$ -<sup>32</sup>P label and an internal fluorescein label, this enzyme removed the radiolabelled  $\gamma$ -phosphate from the 5' end without degrading the transcript (Fig. 1a). No such activity was observed for an RppH mutant with a substitution at an essential active-site residue (E53A)<sup>10</sup>. To demonstrate that RppH removes both the  $\gamma$ - and  $\beta$ -phosphates, we prepared a triphosphorylated RNA substrate (GA(CU)<sub>13</sub>) bearing a single radiolabelled phosphate at either the 5'-terminal  $\alpha$  position or between the first and second nucleotides. Alkaline hydrolysis of either of these substrates

produced pppGp as the major radiolabelled product, as determined by thin-layer chromatography (TLC) (Fig. 2, and Supplementary Fig. S1). After treatment with wild-type RppH, the principal radiolabelled product of alkaline hydrolysis was pGp, as expected for an enzyme that is able to convert triphosphorylated RNA 5' ends to monophosphorylated 5' ends (Fig. 2). Little, if any, ppGp was produced. Treatment with inactive RppH-E53A had no effect. Additional experiments demonstrated that RppH is also active on triphosphorylated RNAs that begin with A, C or U (Supplementary Fig. S2). To ascertain whether this enzyme removes the  $\gamma$ - and  $\beta$ -phosphates in a single step or sequentially, the radiolabelled products generated by



**Figure 1 | RNA pyrophosphohydrolase activity of purified RppH.**

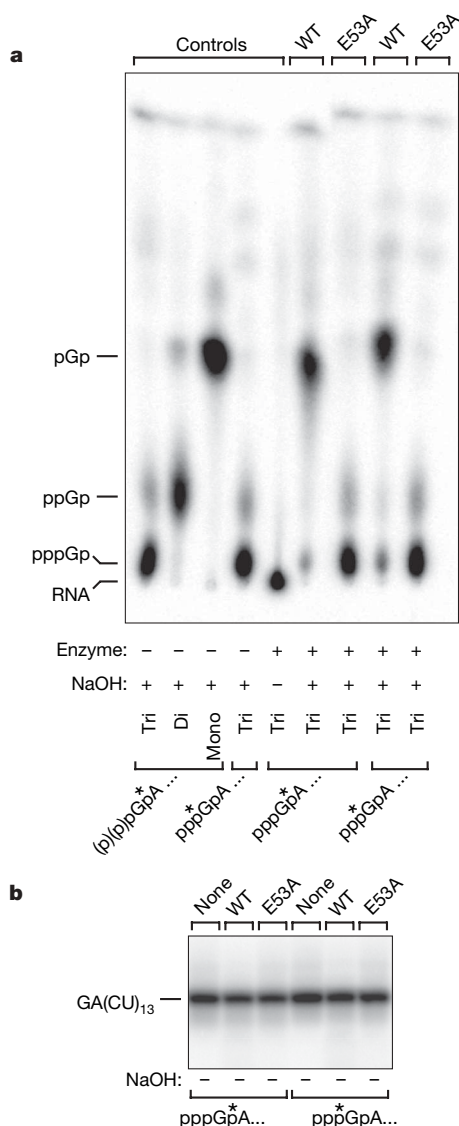
**a**, Electrophoretic assay demonstrating  $\gamma$ -phosphate removal. Triphosphorylated *rpsT* P1 RNA bearing a  $\gamma$ -<sup>32</sup>P label and an internal fluorescein label was treated with purified RppH or RppH-E53A. **b**, TLC assay demonstrating pyrophosphate production. The products of the reaction shown in **a** were analysed by TLC and autoradiography. PP<sub>i</sub>, pyrophosphate; P<sub>i</sub>, orthophosphate. PP<sub>i</sub> (%) = 100 × PP<sub>i</sub>/(PP<sub>i</sub> + P<sub>i</sub>).

<sup>1</sup>Kimmel Center for Biology and Medicine at the Skirball Institute, and Department of Microbiology, New York University School of Medicine, New York, New York 10016, USA.

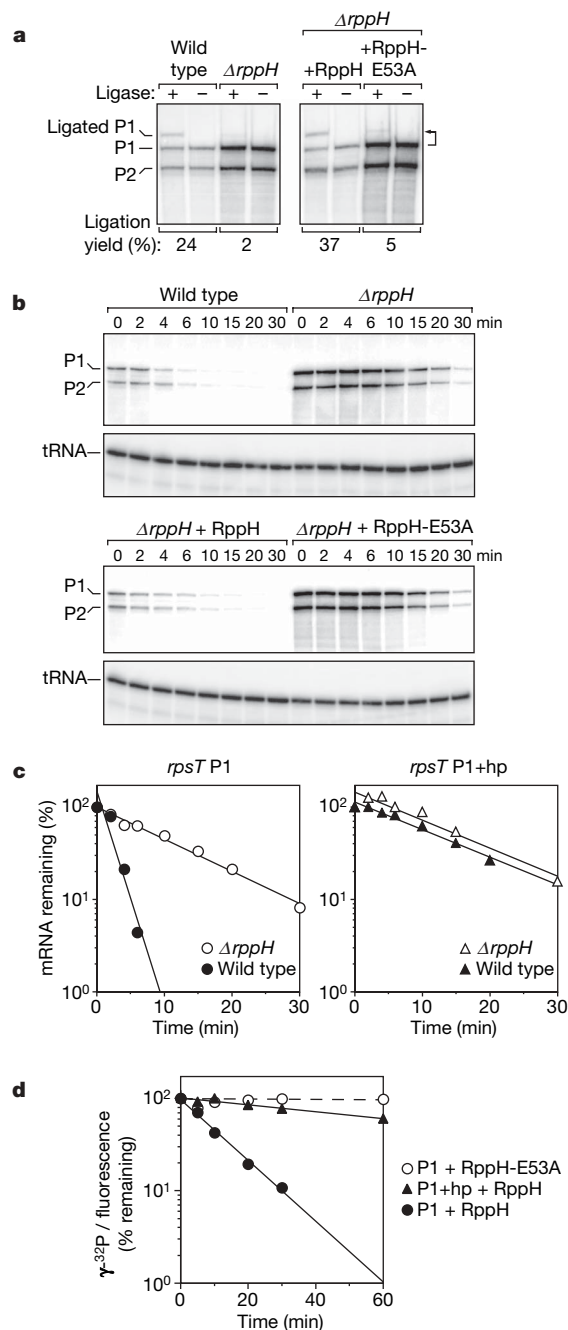
treating the  $\gamma$ - $^{32}\text{P}$  end-labelled transcript with RppH were monitored by TLC as a function of time. Almost all of the radiolabel was released as pyrophosphate, although a small but invariant fraction ( $16 \pm 3\%$ ) was released as orthophosphate (Fig. 1b).

To examine the biological significance of the RNA pyrophosphohydrolase activity of RppH, we tested the effect of a chromosomal *rppH* deletion ( $\Delta rppH$ ) on the 5' phosphorylation state of a transcript of the *E. coli rpsT* gene, which encodes ribosomal protein S20. This gene is transcribed from two promoters to generate a pair of transcripts (P1 and P2) that are degraded by an RNase E-dependent mechanism<sup>11–14</sup>. Previous studies have shown that the *rpsT* P1 transcript can be stabilized by replacing its 5'-terminal triphosphate with a hydroxyl, a finding indicative of a 5'-end-dependent decay mechanism<sup>1</sup>. Consistent with the view that this decay mechanism involves pyrophosphate removal as the initial step, a substantial portion of

*rpsT* P1 mRNA in *E. coli* is monophosphorylated at steady-state. This was judged from an assay (PABLO analysis<sup>1</sup>) in which the 5' phosphorylation state was determined from the ability of monophosphorylated (but not triphosphorylated) 5' ends to undergo splinted ligation with P1-specific DNA oligonucleotides (Fig. 3a). In contrast,



**Figure 2 | Triphosphate-to-monophosphate conversion by purified RppH.** A triphosphorylated transcript ( $\text{GA}(\text{CU})_{13}$ ) bearing a single  $^{32}\text{P}$  label at the 5'-terminal  $\alpha$  position ( $\text{ppp}^*\text{GpA} \dots$ ) or between the first and second nucleotides ( $\text{pppGp}^*\text{A} \dots$ ) was treated with purified RppH (WT) or RppH-E53A (E53A). The radiolabelled products were either subjected to alkaline hydrolysis and then analysed by TLC (a) or examined by gel electrophoresis without hydrolysis to confirm RNA integrity (b). Markers were generated by hydrolysing triphosphorylated (Tri), diphosphorylated (Di) or monophosphorylated (Mono)  $\text{GA}(\text{CU})_{13}$  without prior RppH treatment. (See Supplementary Fig. S1 for a representation of this assay.)



**Figure 3 | RNA pyrophosphohydrolase activity of RppH in *E. coli*.** a, Effect of RppH on the 5' phosphorylation state of the *rpsT* P1 transcript, as determined by PABLO analysis<sup>1</sup> of cellular RNA with P1-specific oligonucleotides. b, Effect of RppH on the decay rate of *rpsT* mRNA, as determined by northern blot analysis of cellular RNA extracted at time intervals after inhibiting transcription. c, RppH-independent decay of *rpsT* mRNA bearing a 5'-terminal stem-loop. The three plasmid-encoded *rpsT* transcripts (P1, P1+hp and P2) were detected by probing for a sequence tag inserted into the 3' untranslated region. Data from representative experiments are shown. d, Inhibition of purified RppH by a 5'-terminal stem-loop. Triphosphorylated *rpsT* P1 and P1+hp RNAs bearing a  $\gamma$ - $^{32}\text{P}$  label and an internal fluorescein label were treated *in vitro* with RppH or RppH-E53A, and representative rates of pyrophosphate removal were plotted. The first three nucleotides of each transcript (AGC) were identical.

**Table 1 | Influence of RppH on selected transcripts in *E. coli***

Transcript	Fold increase in mRNA concentration			mRNA half-life (min)	
	Microarray (E53A/wild type)	Northern blot (E53A/wild type)	Northern blot ( $\Delta rppH/rppH^+$ )	$rppH^+$	$\Delta rppH$
<i>efp</i>	2.9 $\pm$ 0.1	11.3 $\pm$ 1.5	9.8 $\pm$ 0.6	1.6 $\pm$ 0.2	5.3 $\pm$ 0.3
<i>ppa</i>	1.9 $\pm$ 0.1	4.8 $\pm$ 0.1	6.6 $\pm$ 1.4	1.5 $\pm$ 0.2	8.0 $\pm$ 0.3
<i>rpsT</i> P1	2.4 $\pm$ 0.2	7.8 $\pm$ 1.0	5.4 $\pm$ 0.2	1.3 $\pm$ 0.3	6.9 $\pm$ 0.6
<i>rpsT</i> P2	2.4 $\pm$ 0.2	5.3 $\pm$ 0.6	4.3 $\pm$ 0.4	2.2 $\pm$ 0.4	6.9 $\pm$ 1.4
<i>slyB</i>	2.8 $\pm$ 0.2	14.9 $\pm$ 1.0	5.9 $\pm$ 1.2	1.9 $\pm$ 0.1	9.8 $\pm$ 0.4
<i>trxB</i>	3.1 $\pm$ 0.2	4.9 $\pm$ 0.4	7.1 $\pm$ 1.0	2.6 $\pm$ 0.8	28.5 $\pm$ 2.1
<i>yeiP</i>	5.9 $\pm$ 0.5	8.9 $\pm$ 0.7	17.3 $\pm$ 0.4	1.6 $\pm$ 0.1	10.8 $\pm$ 1.0

Transcript concentrations and half-lives were compared in isogenic wild-type ( $rppH^+$ ) and  $\Delta rppH$  strains and in a  $\Delta rppH$  strain complemented with plasmid-encoded wild-type or inactive (E53A) RppH, either by northern blotting or by microarray analysis. The two *rpsT* transcripts were indistinguishable in microarrays. Errors indicate s.d.

few of the P1 transcripts are monophosphorylated in *E. coli* cells lacking RppH. This defect in the  $\Delta rppH$  strain can be fully complemented in trans by a plasmid-borne copy of the wild-type *rppH* gene but not by a mutant allele (*rppH*-E53A). We conclude that RppH is the enzyme principally responsible for pyrophosphate removal from *rpsT* P1 transcripts in *E. coli*.

To investigate the significance of RppH-catalysed pyrophosphate removal for the decay of the *rpsT* P1 transcript, we compared its degradation rate in cells containing or lacking the *rppH* gene. Both the P1 and P2 transcripts were stabilized 3- to 5-fold in the absence of RppH (Fig. 3b), demonstrating the importance of that enzyme for their decay. Rapid turnover was restored in the  $\Delta rppH$  strain by complementation with wild-type RppH but not RppH-E53A. Together, these findings indicate that pyrophosphate removal by RppH triggers rapid degradation of *rpsT* mRNA in *E. coli*.

Previous evidence that *E. coli* transcripts are often stabilized by a 5'-terminal stem-loop<sup>1,4–6,8</sup> suggests that such a structure may exert its influence, at least in part, by hindering pyrophosphate removal by RppH. Consistent with this hypothesis, deletion of the *rppH* gene did not further stabilize an *rpsT* P1 mRNA variant whose lifetime in wild-type cells had been prolonged by adding a 5'-terminal hairpin (P1+hp in Fig. 3c, and Supplementary Fig. S3). To determine whether the RNA pyrophosphohydrolase activity of RppH requires an unpaired 5' end, we compared the ability of purified RppH to remove pyrophosphate from triphosphorylated *rpsT* P1 and P1+hp transcripts. The release of pyrophosphate from the transcript with an unpaired 5' end was nine times faster *in vitro* (Fig. 3d, and Supplementary Fig. S4). Because previous data have shown that RNase E activation by a 5' monophosphate also requires a single-stranded 5' terminus<sup>2</sup>, we conclude that the ability of a 5' stem-loop to stabilize mRNA in *E. coli* is a consequence of both impaired pyrophosphate removal and slow RNase E cleavage caused by sequestration of the 5' end. Whether pyrophosphate removal is also influenced by translating ribosomes remains an open question.

The increased concentration of the *rpsT* transcripts in *E. coli* cells lacking RppH suggested that other targets of this enzyme could be identified by microarray analysis. Triplicate samples of total cellular RNA were isolated from isogenic  $\Delta rppH$  *E. coli* strains complemented by plasmids encoding either wild-type RppH or inactive RppH-E53A and used to probe microarrays representing all the known protein-coding genes of *E. coli* K-12. The abundance of 382 gene transcripts was found to increase significantly (FDR < 0.05) in cells containing RppH-E53A versus wild-type RppH (Supplementary Table S1). As expected, these included *rpsT* mRNA.

To validate that the observed concentration increases were due to impaired mRNA degradation in the absence of active RppH, the longevity and concentration of several of these transcripts were compared in isogenic wild-type ( $rppH^+$ ) and  $\Delta rppH$  *E. coli* strains by northern blotting. In every case, the half-life of the message increased 3- to 11-fold in  $\Delta rppH$  cells, and its steady-state concentration increased 4- to 17-fold (Table 1). That the enhanced longevity of these transcripts in the absence of RppH resulted from impaired pyrophosphate removal was verified for *yeiP* mRNA by showing that its sevenfold greater stability in  $\Delta rppH$  cells was accompanied

by a marked reduction in the percentage of that message that was monophosphorylated (Supplementary Fig. S5). The substantially greater effect of RppH that was measured by blotting versus microarrays suggests that the 382 transcripts shown by gene array analysis to be degraded by an RppH-dependent mechanism may be an underestimate of the actual total.

The half-lives of the transcripts in Table 1 also increased upon RNase E inactivation (1.4- to 3.4-fold; Supplementary Table S2), indicating a role for that endonuclease in degrading the monophosphorylated intermediates produced by RppH. That the absence of RppH caused greater stabilization suggests that those intermediates may each decay by multiple pathways, including some that are independent of RNase E. For example, certain mRNAs sensitive to RppH (such as *yeiP*) are also known targets of RNase G, a minor 5'-monophosphate-dependent RNase E paralogue<sup>15–17</sup>, whereas others might undergo 3' exoribonuclease attack facilitated by 3' oligoadenylate tails added by poly(A) polymerase, another 5'-monophosphate-dependent enzyme<sup>18</sup>. That RppH is not essential for *E. coli* cell growth despite functioning as the master regulator of 5'-end-dependent mRNA decay attests to the availability of alternative, 5'-end-independent degradation pathways.

The biological function of RppH has not previously been defined, even though homologous proteins are widespread among prokaryotic organisms. Genetic experiments with pathogenic bacteria have indicated important roles for this enzyme and its orthologues in invasiveness and virulence<sup>19–22</sup>, which we now suspect may be manifestations of the influence of these proteins on patterns of gene expression. Although studies of purified RppH have shown that it can convert diadenosine oligophosphates into mononucleotides (for example, A[5']pppp[5']A  $\rightarrow$  ATP + AMP)<sup>23</sup>, the biological importance of that *in vitro* activity has not been established. No other catalytic activity of RppH has been reported until now. We therefore propose that this protein (formerly designated NudH/YgdP) and its gene be named RppH to reflect its biological function as an RNA pyrophosphohydrolase.

The ability of RppH to trigger bacterial RNA decay by removing a protective structure at the 5' terminus bears a striking resemblance to the removal of cap structures (m<sup>7</sup>Gppp) from the 5' ends of eukaryotic mRNAs. In each case, a 5'-terminal or 5'-proximal triphosphate is cleaved to produce a monophosphorylated intermediate vulnerable to attack by a 5'-monophosphate-dependent ribonuclease (for example, the endonuclease RNase E in *E. coli* or the 5' exonuclease Xrn1 in eukaryotes)<sup>1,24</sup>. Interestingly, the protein responsible for cap removal in eukaryotic cells (Dcp2) is itself a member of the Nudix family<sup>25,26</sup>. Thus, despite significant structural differences between *E. coli* and human mRNAs, the enzymes that de-protect their 5' termini appear to have evolved from a common ancestor.

## METHODS SUMMARY

**Pyrophosphate release from  $\gamma$ -<sup>32</sup>P-labelled RNA and  $\alpha$ -<sup>32</sup>P-labelled RNA.** Synthetic RNAs were prepared by *in vitro* transcription from a class III  $\phi$ 2.5 T7 promoter<sup>27</sup>, gel-purified and incubated with affinity-purified RppH or RppH-E53A. The products of reactions that contained *rpsT* P1 or P1+hp RNA bearing a 5'-terminal  $\gamma$ -<sup>32</sup>P label and an internal fluorescein label were analysed by electrophoresis on a polyacrylamide-urea gel or by TLC on

PEI-cellulose. The products of reactions that contained triphosphorylated GA(CU)<sub>13</sub>, AG(CU)<sub>13</sub>, CG(A)<sub>26</sub> or UG(A)<sub>26</sub> bearing a single <sup>32</sup>P label at either the 5'-terminal  $\alpha$  position or between the first and second nucleotides were examined both by gel electrophoresis to confirm the integrity of the RNA and by alkaline hydrolysis and TLC to test for pyrophosphate removal. Hydrolysed monophosphorylated and diphosphorylated forms of the same  $\alpha$ -labelled RNAs served as TLC standards.

**Analysis of RNA extracted from *E. coli*.** RNA lifetimes and phosphorylation states were analysed at 37 °C in *E. coli* K-12 strain BW25113 and its isogenic derivative JW2798 $\Delta$ kan, which bears an in-frame deletion of the *rppH* coding region<sup>28</sup>, or at 44 °C in strain TA1025 (*rne*<sup>+</sup>) and its isogenic derivative TA1026<sup>29</sup>, which has a temperature-sensitive RNase E allele (*rne-1*). Total cellular RNA was harvested, as previously described<sup>30</sup>, from cells growing exponentially in MOPS medium containing glucose, uracil and thiamine. In some experiments, isopropyl- $\beta$ -D-thiogalactoside (IPTG) (10  $\mu$ M) was included to induce synthesis of plasmid-encoded RppH. The 5' phosphorylation state of specific transcripts was determined by PABLO analysis, as described<sup>1</sup>. mRNA decay rates were measured after inhibiting transcription with rifampicin. Microarray analysis using *E. coli* Genome 2.0 arrays (Affymetrix) was performed with total cellular RNA extracted from triplicate cultures of JW2798 $\Delta$ kan containing either pPlacRppH or pPlacRppH-E53A and growing exponentially in the presence of IPTG.

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

Received 30 July; accepted 12 November 2007.

- Celesnik, H., Deana, A. & Belasco, J. G. Initiation of RNA decay in *Escherichia coli* by 5' pyrophosphate removal. *Mol. Cell* **27**, 79–90 (2007).
- Mackie, G. A. Ribonuclease E is a 5'-end-dependent endonuclease. *Nature* **395**, 720–723 (1998).
- Jiang, X. & Belasco, J. G. Catalytic activation of multimeric RNase E and RNase G by 5'-monophosphorylated RNA. *Proc. Natl Acad. Sci. USA* **101**, 9211–9216 (2004).
- Emory, S. A., Bouvet, P. & Belasco, J. G. A 5'-terminal stem-loop structure can stabilize mRNA in *Escherichia coli*. *Genes Dev.* **6**, 135–148 (1992).
- Bouvet, P. & Belasco, J. G. Control of RNase E-mediated RNA degradation by 5'-terminal base pairing in *E. coli*. *Nature* **360**, 488–491 (1992).
- Bricker, A. L. & Belasco, J. G. Importance of a 5' stem-loop for longevity of *papA* mRNA in *Escherichia coli*. *J. Bacteriol.* **181**, 3587–3590 (1999).
- Mackie, G. A. Stabilization of circular *rpsT* mRNA demonstrates the 5'-end dependence of RNase E action in vivo. *J. Biol. Chem.* **275**, 25069–25072 (2000).
- Baker, K. E. & Mackie, G. A. Ectopic RNase E sites promote bypass of 5'-end-dependent mRNA decay in *Escherichia coli*. *Mol. Microbiol.* **47**, 75–88 (2003).
- McLennan, A. G. The Nudix hydrolase superfamily. *Cell. Mol. Life Sci.* **63**, 123–143 (2006).
- Mildvan, A. S. *et al.* Structures and mechanisms of Nudix hydrolases. *Arch. Biochem. Biophys.* **433**, 129–143 (2005).
- Mackie, G. A. & Parsons, G. D. Tandem promoters in the gene for ribosomal protein S20. *J. Biol. Chem.* **258**, 7840–7846 (1983).
- Mackie, G. Specific endonucleolytic cleavage of the mRNA for ribosomal protein S20 of *Escherichia coli* requires the product of the *ams* gene in vivo and in vitro. *J. Bacteriol.* **173**, 2488–2497 (1991).
- Mackie, G. A. Secondary structure of the mRNA for ribosomal protein S20. Implications for cleavage by ribonuclease E. *J. Biol. Chem.* **267**, 1054–1061 (1992).
- Ow, M. C., Perwez, T. & Kushner, S. R. RNase G of *Escherichia coli* exhibits only limited functional overlap with its essential homologue, RNase E. *Mol. Microbiol.* **49**, 607–622 (2003).
- Tock, M. R., Walsh, A. P., Carroll, G. & McDowall, K. J. The CafA protein required for the 5'-maturation of 16 S rRNA is a 5'-end-dependent ribonuclease that has context-dependent broad sequence specificity. *J. Biol. Chem.* **275**, 8726–8732 (2000).
- Jiang, X., Diwa, A. & Belasco, J. G. Regions of RNase E important for 5'-end-dependent RNA cleavage and autoregulated synthesis. *J. Bacteriol.* **182**, 2468–2475 (2000).
- Lee, K., Bernstein, J. A. & Cohen, S. N. RNase G complementation of *rne* null mutation identifies functional interrelationships with RNase E in *Escherichia coli*. *Mol. Microbiol.* **43**, 1445–1456 (2002).
- Feng, Y. & Cohen, S. N. Unpaired terminal nucleotides and 5' monophosphorylation govern 3' polyadenylation by *Escherichia coli* poly(A) polymerase I. *Proc. Natl Acad. Sci. USA* **97**, 6415–6420 (2000).
- Mitchell, S. J. & Minnick, M. F. Characterization of a two-gene locus from *Bartonella bacilliformis* associated with the ability to invade human erythrocytes. *Infect. Immun.* **63**, 1552–1562 (1995).
- Badger, J. L., Wass, C. A. & Kim, K. S. Identification of *Escherichia coli* K1 genes contributing to human brain microvascular endothelial cell invasion by differential fluorescence induction. *Mol. Microbiol.* **36**, 174–182 (2000).
- Ismail, T. M., Hart, C. A. & McLennan, A. G. Regulation of dinucleoside polyphosphate pools by the YgdP and ApaH hydrolases is essential for the ability of *Salmonella enterica* serovar typhimurium to invade cultured mammalian cells. *J. Biol. Chem.* **278**, 32602–32607 (2003).
- Edelstein, P. H. *et al.* *Legionella pneumophila* NudA Is a Nudix hydrolase and virulence factor. *Infect. Immun.* **73**, 6567–6576 (2005).
- Bessman, M. J. *et al.* The gene *ygdP*, associated with the invasiveness of *Escherichia coli* K1, designates a Nudix hydrolase, Orf176, active on adenosine (5')-pentaphospho-(5')-adenosine (Ap5A). *J. Biol. Chem.* **276**, 37834–37838 (2001).
- Muhlrad, D., Decker, C. J. & Parker, R. Deadenylation of the unstable mRNA encoded by the yeast *MFA2* gene leads to decapping followed by 5'→3' digestion of the transcript. *Genes Dev.* **8**, 855–866 (1994).
- Dunkley, T. & Parker, R. The DCP2 protein is required for mRNA decapping in *Saccharomyces cerevisiae* and contains a functional MutT motif. *EMBO J.* **18**, 5411–5422 (1999).
- Wang, Z., Jiao, X., Carr-Schmid, A. & Kiledjian, M. The hDcp2 protein is a mammalian mRNA decapping enzyme. *Proc. Natl Acad. Sci. USA* **99**, 12663–12668 (2002).
- Coleman, T. M., Wang, G. & Huang, F. Superior 5' homogeneity of RNA from ATP-initiated transcription under the T7  $\phi$ 2.5 promoter. *Nucleic Acids Res.* **32**, e14 (2004).
- Baba, T. *et al.* Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).
- Arnold, T. E., Yu, J. & Belasco, J. G. mRNA stabilization by the *ompA* 5' untranslated region: two protective elements hinder distinct pathways for mRNA degradation. *RNA* **4**, 319–330 (1998).
- Emory, S. A. & Belasco, J. G. The *ompA* 5' untranslated RNA segment functions in *Escherichia coli* as a growth-rate-regulated mRNA stabilizer whose activity is unrelated to translational efficiency. *J. Bacteriol.* **172**, 4472–4481 (1990).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We are grateful to D. Guttman for his assistance in the discovery that purified RppH has RNA pyrophosphohydrolase activity. This research was supported by a grant to J.G.B. from the National Institutes of Health.

**Author Contributions** A.D., H.C. and J.G.B. planned the studies, interpreted the data and wrote the manuscript. A.D. and H.C. performed the experiments.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence and requests for materials should be addressed to J.G.B. ([belasco@saturn.med.nyu.edu](mailto:belasco@saturn.med.nyu.edu)).

## METHODS

**Plasmids.** Plasmid pET-RPST1, when linearized with *NotI*, allows *in vitro* synthesis of an RNA identical to the *rpsT* P1 transcript but for a U → G substitution at the second nucleotide<sup>1</sup>. This substitution facilitates transcription by T7 RNA polymerase from a class III  $\phi$ 2.5 T7 promoter that efficiently produces A-initiated transcripts<sup>27</sup>. Plasmid pET-RPST1 + hp was constructed from pET-RPST1 by inserting the sequence AGCGCCGCTCGAGCGGCGCT at the beginning of the transcribed region. Plasmid pRPST1 contains a complete copy of the *E. coli rpsT* gene along with a unique sequence tag inserted into the 3' untranslated region (UTR)<sup>1</sup>. Plasmid pRPST1 + hp is a derivative of pRPST1 in which the *rpsT* P1 promoter has been replaced by a *bla* promoter and an inverted repeat (ATCGCCGCTCGAGCGGCGAT) has been added to the 5' end of the P1 transcriptional unit. Plasmid pPlacRppH6 is a pRNGL3<sup>17</sup> derivative that encodes amino-terminally hexahistidine-tagged RppH under the control of an IPTG-inducible *lacUV5* promoter. pPlacRppH6-E53A was constructed from pPlacRppH6 by a codon substitution (GAA → GCA) at position 53. Removal of the six histidine codons generated plasmids pPlacRppH and pPlacRppH-E53A, which were used to test complementation of the *ΔrppH* phenotype of JW2798*Akan*.

**Affinity purification of RppH.** A 1–2 litre culture of *E. coli* JW2798*Akan* containing pPlacRppH6 or pPlacRppH6-E53A was induced with IPTG (1 mM) for 4 h and harvested. The bacterial pellets were resuspended in 10 ml of buffer E (10 mM HEPES pH 7.6, 300 mM NaCl, 0.25% Genapol, 0.1 mM PMSF) containing protease inhibitors (Complete, EDTA-free; Roche) and lysed by passage through a French press. The lysate was treated for 1 h with DNase I (0.1 mg; Roche) in the presence of MgSO<sub>4</sub> (20 mM) and cleared by centrifugation at 14,500g for 30 min. The hexahistidine-tagged RppH protein was then attached to TALON beads (1–2 ml; BD Biosciences) by incubation for 1 h at 4 °C, washed with buffer E supplemented with protease inhibitors and containing 20 mM imidazole, and eluted with buffer E supplemented with protease inhibitors and containing 250 mM imidazole (pH 7.6). After overnight dialysis against buffer E, the protein was purified a second time with TALON beads, concentrated by centrifugal ultrafiltration, and stored at –20 °C in buffer E containing 25% (by volume) glycerol.

**Detection of pyrophosphate release from  $\gamma$ -<sup>32</sup>P-labelled RNA.** Doubly labelled *rpsT* P1 or P1 + hp RNA bearing a 5'-terminal  $\gamma$ -<sup>32</sup>P label and an internal fluorescein label was synthesized by *in vitro* transcription in a mixture (40  $\mu$ l) containing Tris-Cl (40 mM, pH 7.9), MgCl<sub>2</sub> (6 mM), NaCl (10 mM), dithiothreitol (10 mM), spermidine (2 mM), GTP (1 mM), CTP (1 mM), UTP (1 mM), ATP (0.5 mM), fluorescein-UTP (0.25 mM; Roche), [ $\gamma$ -<sup>32</sup>P]ATP (50  $\mu$ Ci, 0.28  $\mu$ M), RNasin (20 units; Promega), plasmid pET-RPST1 or pET-RPST1 + hp linearized by *NotI* cleavage (0.13 pmol), and T7 RNA polymerase (40 units; New England Biolabs). (These DNA templates each contained a class III  $\phi$ 2.5 T7 promoter for efficient production of A-initiated transcripts<sup>27</sup>.) After incubation for 4 h at 37 °C, the RNA was gel-purified.

The doubly labelled RNA (16,000 c.p.m., 18.2 nM) was incubated with purified RppH (75 nM) in a solution (160  $\mu$ l) containing HEPES (20 mM, pH 7.6), MgCl<sub>2</sub> (5 mM), dithiothreitol (1 mM) and glycerol (1%) for 0–60 min at 37 °C. Reaction samples (20  $\mu$ l) were quenched at time intervals with 5  $\mu$ l of EDTA (10 mM, pH 8.0) and analysed by electrophoresis on a 6% polyacrylamide–8 M urea gel or by TLC on PEI-cellulose (J. T. Baker) developed with potassium phosphate buffer (0.3 M, pH 7.5). Band or spot intensities were quantified and compared by using a Molecular Dynamics Storm 820 PhosphorImager or a Molecular Dynamics FluorImager 575 and ImageQuant software.

**Detection of pyrophosphate release from  $\alpha$ -<sup>32</sup>P-labelled RNA.** Triphosphorylated GA(CU)<sub>13</sub>, AG(CU)<sub>13</sub>, CG(A)<sub>26</sub> and UG(A)<sub>26</sub> oligoribonucleotides bearing a single <sup>32</sup>P label at either the 5'-terminal  $\alpha$  position or between the first and second nucleotides were synthesized for 6–8 h at 37 °C in a mixture (40  $\mu$ l) containing Tris-Cl (40 mM, pH 7.9), MgCl<sub>2</sub> (6 mM), spermidine (2 mM), dithiothreitol (20 mM), rRNasin (40 units, Promega), T7 RNA polymerase (100 units; New England Biolabs), Ampliscribe T7-Flash enzyme solution (2  $\mu$ l; Epicentre), the appropriate DNA template bearing a class III  $\phi$ 2.5 T7 promoter<sup>27</sup> (30 pmol), GTP (1 mM unlabelled or 50  $\mu$ Ci = 0.42  $\mu$ M  $\alpha$ -<sup>32</sup>P-labelled), ATP (1 mM unlabelled or 50  $\mu$ Ci = 0.42  $\mu$ M  $\alpha$ -<sup>32</sup>P-labelled), CTP (1 mM unlabelled or 50  $\mu$ Ci = 0.42  $\mu$ M  $\alpha$ -<sup>32</sup>P-labelled) and UTP (1 mM unlabelled). Monophosphorylated and diphosphorylated forms of the same RNAs were synthesized by replacing the 5'-terminal NTP in the transcription reaction with either NMP or NDP (10 mM). Each RNA was gel-purified.

The  $\alpha$ -labelled RNA (40,000 c.p.m., 0.3 nM) was incubated with purified RppH (75 nM) in a solution (20  $\mu$ l) containing HEPES (20 mM, pH 7.6),

MgCl<sub>2</sub> (5 mM), dithiothreitol (1 mM) and glycerol (1%) for 2 h at 37 °C. The reaction was stopped by adding 5  $\mu$ l of EDTA (100 mM, pH 8.0). To confirm the integrity of the RNA, 15  $\mu$ l of the reaction product was examined by electrophoresis on a 16% polyacrylamide–8 M urea gel. The remaining 10  $\mu$ l was subjected to alkaline hydrolysis by incubation with 5  $\mu$ l of NaOH (0.2 M) at 95 °C for 15 min. After neutralization with 5  $\mu$ l of formic acid (3 M), the reaction products were analysed by TLC on a PEI-cellulose plate (J. T. Baker) developed with potassium phosphate buffer (0.3 M, pH 3.3), and radiolabelled products were detected with a Molecular Dynamics Storm 820 PhosphorImager.

**RNA extraction from *E. coli*.** Measurements of RNA lifetime and phosphorylation state were performed in *E. coli* K-12 strain BW25113<sup>31</sup> and its isogenic derivative JW2798*Akan*, which bears an in-frame deletion of all but the first codon and last six codons of the *rppH* coding region, or in strain TA1025 (*rne*<sup>+</sup>) and its isogenic derivative TA1026<sup>29</sup>, which has a temperature-sensitive RNase E allele (*rne*-1). JW2798*Akan* was constructed by excising the *kan* cassette that had replaced the *rppH* gene of JW2798<sup>28</sup>, a Keio strain provided by the National Institute of Genetics (Japan).

Total cellular RNA was harvested, as previously described<sup>30</sup>, from *E. coli* growing exponentially at 37 °C in MOPS medium containing glucose (0.2%), uracil (20  $\mu$ g ml<sup>–1</sup>) and thiamine (1  $\mu$ g ml<sup>–1</sup>), or 10 min after a temperature increase from 30 to 44 °C in the case of TA1025 and TA1026. In some experiments, the culture medium also contained IPTG (10  $\mu$ M) to induce synthesis of plasmid-encoded RppH.

**PABLO analysis.** The 5' phosphorylation state of RNA in *E. coli* was determined by PABLO analysis, as described<sup>1</sup>. The oligonucleotides used in these experiments were Y<sub>rpsT-P1</sub> (5'-ACTCGTTACGTAGTGATCAAGTTATCATTCATATTGTC-3'), Y<sub>yeiP</sub> (5'-AGTCGAAAATGTCAAAAATATCAAGTTATCATTCATATTGTC-3') and X<sub>90</sub> (5'-CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCGAACAATATGAATGATACTTG-3'). The 5' end of *yeiP* mRNA (5'-AUUUUUUG ...) maps to a site 35 base pairs (bp) upstream of the initiation codon and 7 bp downstream of the *yeiP* promoter (TTGCCC–17 bp–TACTTT), whose identity was confirmed by mutation. The radioactive probes used to detect *rpsT* and *yeiP* mRNA were a 5'-end-labelled oligonucleotide complementary to a 5' UTR segment shared by the *rpsT* P1 and P2 transcripts (5'-GTCCAACCTCCC-AAATGTGTTTC-3') and an internally labelled probe generated by random priming (High Prime, Roche) on *yeiP* DNA.

**Measurements of RNA half-life.** Total cellular RNA was extracted from *E. coli* at time intervals after inhibiting transcription with rifampicin (0.2 mg ml<sup>–1</sup>), and equal amounts (10  $\mu$ g) were subjected to gel electrophoresis on polyacrylamide (6% or 4.5%) containing 8 M urea. The RNA was transferred to a Hybond-XL membrane (Amersham) by electroblotting and ultraviolet crosslinking, and probed with an internally radiolabelled probe generated by random priming (High Prime, Roche) on the coding region of *yeiP*, *ppa*, *efp*, *slyB* or *trxB*, or with a 5'-radiolabelled oligonucleotide probe complementary to a 5' UTR segment shared by the *rpsT* P1 and P2 transcripts (5'-GTCCAACCTCCC-AAATGTGTTTC-3'), a sequence tag inserted into the 3' UTR of *rpsT* (5'-CAAAGATCGGGG-TGGGGGTCTAAG-3'), an internal region of *yeiP* (5'-GCCGTTGTAATTCA-GTACCA-3'), an internal region of *ppa* (5'-TCTGCGTTAGCCGGGATCTC-3'), 5S rRNA (5'-ACTACCATCGGCGCTACGGC-3') or *cysT* tRNA (5'-GGAGTCGAACCGGACTAGACGG-3'). Radioactive bands were visualized with a Molecular Dynamics Storm 820 PhosphorImager, and band intensities were quantified by using ImageQuant software. RNA half-lives were calculated by linear regression analysis of data, which were obtained from at least two or three independent experiments.

**Microarray analysis.** Total cellular RNA was extracted from triplicate cultures of JW2798*Akan* containing either pPlacRppH or pPlacRppH-E53A and growing exponentially in the presence of IPTG (10  $\mu$ M) to induce RppH synthesis. Complementary DNA was prepared by random-primed reverse transcription with SuperScript II (Invitrogen), fragmented with DNase I (GE Biosciences), biotin-labelled with the GeneChip DNA labelling reagent (Affymetrix) and terminal deoxynucleotidyl transferase (Promega), and used to probe *E. coli* Genome 2.0 arrays (Affymetrix). The microarrays were scanned with an Affymetrix GeneChip Scanner 3000, and the raw data were scaled and quantified with Affymetrix GCOS software. Calculations of relative mRNA concentrations, including normalization, were performed with the MAS 5 algorithm.

31. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97**, 6640–6645 (2000).

# Translational control of intron splicing in eukaryotes

Olivier Jaillon<sup>1,2,3\*</sup>, Khaled Bouhouche<sup>4,5,6,7,8\*</sup>, Jean-François Gout<sup>9</sup>, Jean-Marc Aury<sup>1,2,3</sup>, Benjamin Noel<sup>1,2,3</sup>, Baptiste Saudemont<sup>4,5</sup>, Mariusz Nowacki<sup>4,5</sup>, Vincent Serrano<sup>4,5</sup>, Betina M. Porcel<sup>1,2,3</sup>, Béatrice Ségurens<sup>1</sup>, Anne Le Mouél<sup>4,5</sup>, Gersende Lepère<sup>4,5</sup>, Vincent Schächter<sup>1,2,3</sup>, Mireille Bétermier<sup>6,7,8</sup>, Jean Cohen<sup>6,7,8</sup>, Patrick Wincker<sup>1,2,3</sup>, Linda Sperling<sup>6,7,8</sup>, Laurent Duret<sup>9</sup> & Eric Meyer<sup>4,5</sup>

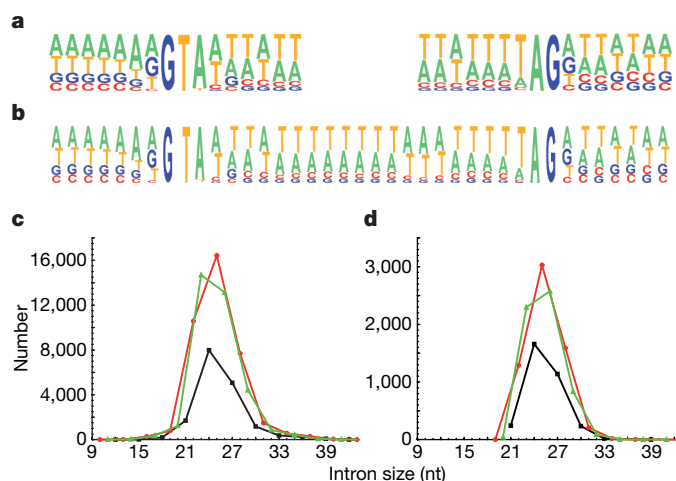
Most eukaryotic genes are interrupted by non-coding introns that must be accurately removed from pre-messenger RNAs to produce translatable mRNAs<sup>1</sup>. Splicing is guided locally by short conserved sequences, but genes typically contain many potential splice sites, and the mechanisms specifying the correct sites remain poorly understood. In most organisms, short introns recognized by the intron definition mechanism<sup>2</sup> cannot be efficiently predicted solely on the basis of sequence motifs<sup>3</sup>. In multicellular eukaryotes, long introns are recognized through exon definition<sup>2</sup> and most genes produce multiple mRNA variants through alternative splicing<sup>4</sup>. The nonsense-mediated mRNA decay<sup>5,6</sup> (NMD) pathway may further shape the observed sets of variants by selectively degrading those containing premature termination codons, which are frequently produced in mammals<sup>7,8</sup>. Here we show that the tiny introns of the ciliate *Paramecium tetraurelia* are under strong selective pressure to cause premature termination of mRNA translation in the event of intron retention, and that the same bias is observed among the short introns of plants, fungi and animals. By knocking down the two *P. tetraurelia* genes encoding UPF1, a protein that is crucial in NMD, we show that the intrinsic efficiency of splicing varies widely among introns and that NMD activity can significantly reduce the fraction of unspliced mRNAs. The results suggest that, independently of alternative splicing, species with large intron numbers universally rely on NMD to compensate for suboptimal splicing efficiency and accuracy.

With an average length of 25 nucleotides (nt), the spliceosomal introns of *P. tetraurelia* are among the shortest reported in any eukaryote<sup>9</sup>. Annotation of the somatic genome<sup>10</sup>, which was based in part on the alignment of 78,110 expressed sequence tags (ESTs), predicted a total of 39,642 protein-coding genes containing 90,282 introns (2.3 introns per gene on average), 96.8% of which are between 20 and 34 nt in length. That such small introns are recognized through intron definition, as in other unicellular eukaryotes<sup>11</sup>, is supported by our observation that introns inserted in the coding sequence of a green fluorescent protein reporter are efficiently spliced out (not shown). Alternative splicing is very limited: not a single case of exon skipping was observed, and fewer than 0.9% of the 13,498 introns covered by at least two ESTs were found to use alternative splice sites, usually closely spaced 3' sites (results not shown). The compositional profiles of 5' and 3' splice sites revealed that only the first and last three bases of introns are highly constrained (Fig. 1); by comparison with short introns of other eukaryotes<sup>3</sup>, these profiles seem to have a very low information content.

The size distribution of predicted introns shows a conspicuous deficit in introns whose length is a multiple of 3 (hereafter called

3*n* introns): these represent only 18.7% of the total, in contrast with 42.3% and 39.0% for 3*n* + 1 and 3*n* + 2 introns, respectively (Fig. 1c). Because intron prediction relies heavily on the reconstruction of open reading frames and is therefore more likely to overlook short 3*n* introns that do not contain in-frame stop codons, we extracted a high-confidence data set by selecting 6,137 gene models for which each of the predicted introns was confirmed by the alignment of at least one EST. Among the 15,286 confirmed introns, 3*n* introns are still strongly under-represented (Fig. 1d): 21.6% of the total, in contrast with 40.2% and 38.2% for 3*n* + 1 and 3*n* + 2 introns, respectively (significantly different from a random distribution;  $\chi^2 = 956$ ,  $P < 10^{-16}$ ). Thus, the under-representation of 3*n* introns is not attributable to annotation artefacts.

One particular feature of 3*n* introns is that they would not cause a frame shift during the translation of intron-retaining mRNAs, whereas the retention of most 3*n* + 1 or 3*n* + 2 introns (93.8% and 84.0% of those in the confirmed set, respectively) would introduce a premature termination codon (PTC) in the downstream exons. To



**Figure 1 | Characteristics of *P. tetraurelia* introns.** **a**, Compositional profiles of the 5' (left) and 3' (right) splice sites, including seven nucleotides outside and nine nucleotides inside the intron ( $n = 15,286$  EST-confirmed introns). **b**, Compositional profile of the entire length of 25-nt introns (the most abundant size class), with seven nucleotides of the flanking exons on both sides ( $n = 3,028$  EST-confirmed introns). **c**, Size distribution of the 90,282 annotated introns. 3*n*, 3*n* + 1 and 3*n* + 2 introns are shown in black, red and green, respectively. **d**, Size distribution of the 15,286 EST-confirmed introns.

<sup>1</sup>Genoscope (CEA), 2 rue Gaston Crémieux CP5706, 91057 Evry, France. <sup>2</sup>CNRS, UMR 8030, 2 rue Gaston Crémieux CP5706, 91057 Evry, France. <sup>3</sup>Université d'Evry, 91057 Evry, France. <sup>4</sup>École Normale Supérieure, Laboratoire de Génétique Moléculaire, 46 rue d'Ulm, 75005 Paris, France. <sup>5</sup>CNRS, UMR 8541, 46 rue d'Ulm, 75005 Paris, France. <sup>6</sup>CNRS, Centre de Génétique Moléculaire, UPR 2167, 91198 Gif-sur-Yvette, France. <sup>7</sup>Université Paris-Sud, 91405 Orsay, France. <sup>8</sup>Université Pierre et Marie Curie – Paris 6, 75005 Paris, France. <sup>9</sup>CNRS, Laboratoire de Biométrie et Biologie Évolutive, UMR 5558, Université de Lyon, Université Lyon 1, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne, France.

\*These authors contributed equally to this work.

confirm a possible link with translation, size distributions were plotted separately for introns that do or do not contain an in-frame UGA, the only stop codon used in *Paramecium* (Fig. 2). Strikingly, the fraction of  $3n$  introns is only 19.1% in the stopless subset, but close to the expected one-third in the stop-containing subset (35.7%). As a consequence of the larger size of the stopless subset, in-frame UGAs are about twice as frequent in the whole set of  $3n$  introns as in other size classes (Supplementary Table 1 and Supplementary Figs 1 and 2).

The specific counter-selection of stopless  $3n$  introns suggests that *Paramecium* introns are under strong selective pressure to cause premature translation termination in the event of intron retention. A similar bias would easily have been overlooked in other eukaryotes that have longer introns and use three stop codons, because most introns are expected to contain in-frame stops. We therefore examined separately the stopless and stop-containing subsets of complementary-DNA-confirmed introns from *Arabidopsis thaliana*, *Homo sapiens*, *Caenorhabditis elegans* and *Drosophila melanogaster* (Fig. 3 and Supplementary Fig. 3). In all species a highly statistically significant deficit in  $3n$  introns is observed among stopless introns but not among stop-containing introns ( $P < 10^{-12}$ ; Supplementary Table 2). The bias is observed only for short introns, suggesting that it may apply to those recognized by intron definition (Supplementary Table 2). In *Schizosaccharomyces pombe*, whose introns are all recognized by intron definition<sup>11</sup>, the bias is obvious among annotated introns (Supplementary Fig. 3), and the same trend is observed in a small cDNA-confirmed subset (Supplementary Table 2). Thus, stopless  $3n$  introns recognized through intron definition seem to be counter-selected in all intron-rich eukaryotic genomes.

The *P. tetraurelia* genome offers insight into the evolution of intron sequences, as the result of a well-preserved whole-genome duplication that has allowed the identification of 12,026 pairs of duplicated genes<sup>10</sup>. Alignment of the 1,112 pairs belonging to the EST-confirmed set revealed only a handful of cases of intron gains or losses and showed that in at least 37% of 2,774 intron pairs, at least one intron has changed size class since the duplication. The selective pressure that maintains  $3n$  depletion in the face of such length variation must therefore be quite strong. In addition, 6,443 pairs of introns of identical sizes provide evidence for evolutionary conservation of stop codons in  $3n$  introns. Indeed, 59% of in-frame UGAs in  $3n$  introns are conserved in the duplicate, in contrast with 38% for out-of-frame UGAs in  $3n$  introns and 37% for in-frame UGAs in non- $3n$  introns ( $P < 0.001$ ; see Supplementary Fig. 4).

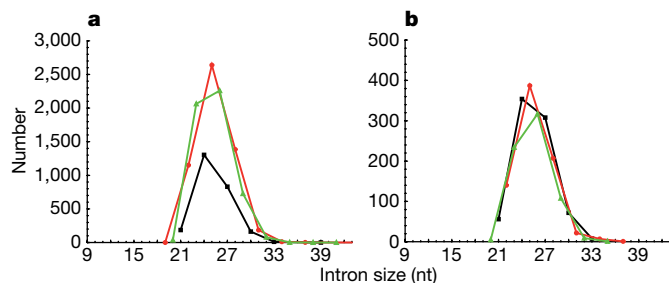
Because no mechanism other than translation itself is currently known to recognize in-frame stop codons, the finding that eukaryotic short introns are under strong selective pressure to introduce PTCs implies that these introns are translated at a substantial frequency. If translation occurs only in the cytoplasm, this further implies that introns are frequently retained in exported mRNAs, which could be linked to the weakness of splicing signals. During the pioneer round of translation<sup>12</sup>, the PTCs resulting from intron retention will trigger mRNA degradation by NMD, thereby protecting cells from

possible dominant-negative effects of truncated proteins. Relying on NMD to compensate for inefficient splicing would make stopless  $3n$  introns dangerous because their retention, which does not introduce any PTC, can still affect protein function.

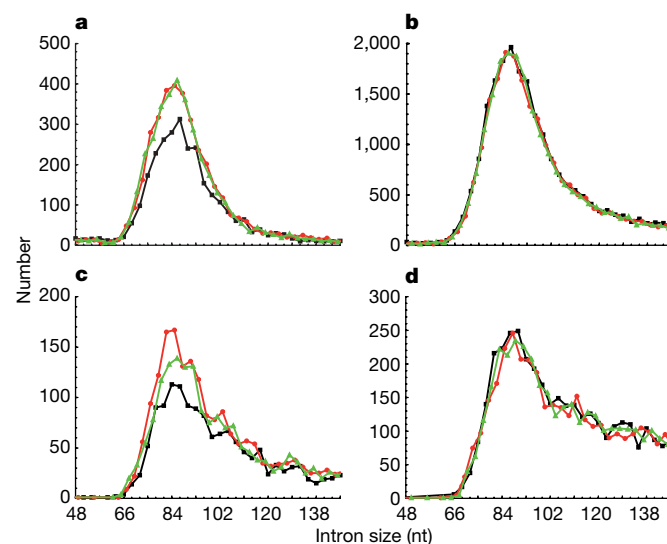
As a first test of these hypotheses, we used the double-stranded RNA feeding technique<sup>13</sup> to knock down NMD activity in *P. tetraurelia*. Targeting either or both of the two *UPF1* paralogs consistently resulted in a modest but significant decrease in *UPF1* mRNA levels (more than twofold; Supplementary Fig. 5). This treatment reduced vegetative growth rate by about 30% and completely blocked meiosis (not shown). We then used an oligo(dT)-primed RT-PCR assay to monitor the fraction of unspliced mRNAs for different types of introns, focusing on introns that were found to be maintained in some ESTs or that had non-consensus bases at the third or third-before-last positions (Supplementary Table 3). Spliced and unspliced versions were amplified together in the same PCR reaction with primers flanking the introns, resolved by electrophoresis and quantified (Fig. 4). Even in normal NMD conditions, a variable fraction of unspliced mRNAs was detected for most of the  $3n + 1$ ,  $3n + 2$  or stop-containing  $3n$  introns tested. Knocking down *UPF1* genes increased this fraction by 10–588% (Fig. 4 and Supplementary Fig. 6). Thus, splicing efficiency varies widely among these introns, and NMD can efficiently reduce the unspliced fraction, at least for some of them.

In contrast, all three stopless  $3n$  introns tested seem to be very efficiently spliced: only intron 7 showed a small but detectable fraction of unspliced mRNAs, and as expected this was not altered by *UPF1* knockdown. This suggests that many of the stopless  $3n$  introns present in the genome are tolerated because they happen to be so efficiently spliced that translational control of splicing is not required. In support of this idea, the analysis of introns occasionally retained in ESTs from wild-type cells shows that the retention rate of stopless introns is significantly lower for  $3n$  introns than for  $3n + 1$  or  $3n + 2$  introns (0.55%, in contrast with 0.86% or 0.79%; see Supplementary Table 4). On average, stopless  $3n$  introns also have stronger splicing signals than other types of introns (Supplementary Table 5).

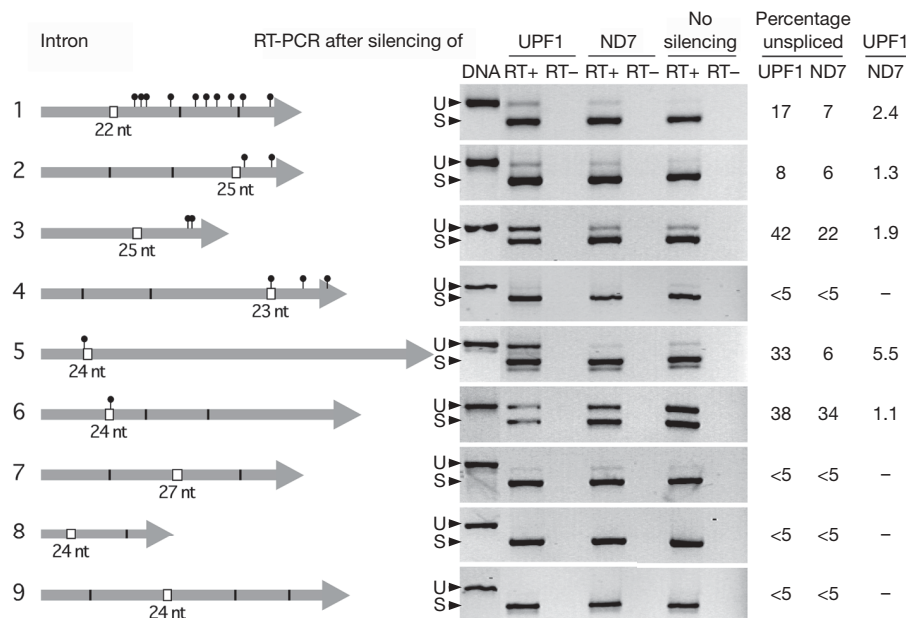
The prominent role of NMD in shaping the observed bias is further supported by knockdown of the *Paramecium* *UPF2* gene (Supplementary Fig. 6) by RNA-mediated interference (RNAi), and by an analysis of the last introns of genes across species. Mammals are



**Figure 2 | Size distributions of the 13,050 stopless and 2,236 stop-containing introns from the EST-confirmed set.** **a**, Stopless introns; **b**, stop-containing introns.  $3n$ ,  $3n + 1$  and  $3n + 2$  introns are shown in black, red and green, respectively.



**Figure 3 | Size distributions of introns in other eukaryotes.** The graphs show the lower modes of the distributions of stopless (**a**, **c**) and stop-containing (**b**, **d**) confirmed introns from *A. thaliana* (**a**, **b**;  $n = 10,482$  and  $87,440$ , respectively) and *H. sapiens* (**c**, **d**;  $n = 6,835$  and  $123,915$ , respectively).  $3n$ ,  $3n + 1$  and  $3n + 2$  introns are shown in black, red and green, respectively.



**Figure 4 | Accumulation of unspliced mRNAs after *UPF1* knockdown.** Tested introns (boxes), their positions within coding sequences (thick grey arrows; see Supplementary Table 3), PTCs introduced by their retention (black pinheads) and exon–exon junctions resulting from the splicing of other introns (vertical black lines) are shown schematically. RT–PCR

peculiar in that a PTC will trigger NMD only if it is located at least 50 nt upstream of the last exon–exon junction<sup>5,6</sup>. The retention of the last intron of a gene therefore cannot be detected by NMD in mammals, even if it introduces a PTC. Accordingly, stopless *3n* introns are not under-represented among the last introns of genes in *H. sapiens*, whereas they are in non-mammalian species (Supplementary Table 2).

It has been proposed that the appearance of genome-wide alternative splicing during the evolution of multicellular organisms was linked to the weakening of strong splice sites of ancestral introns recognized through intron definition<sup>4</sup>. We found that alternative splicing does not occur to any significant degree in *Paramecium*, an organism in which a relatively large intron number is associated with very weak splice sites and with the strong counter-selection of those introns that cannot be detected by NMD. The finding that the latter features are common to various intron-rich eukaryotes suggests that, independently of alternative splicing, it may be more advantageous to rely on NMD surveillance than to evolve a more efficient splicing system. Supporting this view, the rare species that seem to have lost NMD are almost entirely devoid of introns<sup>14</sup>.

Finally, we note that the observed bias is also compatible with the controversial proposal, based on studies of the nonsense-associated alternative splicing<sup>15–17</sup> and suppression of splicing<sup>18,19</sup> effects, that the translatability of pre-mRNA sequences can influence splice site choice<sup>20,21</sup>. Although this idea was revived by the finding that a substantial fraction of mammalian NMD events occurs in the nucleus<sup>22,23</sup> and by the controversial possibility of nuclear translation<sup>24–27</sup>, it should be emphasized that it does not necessarily imply nuclear translation before splicing. RNA interference can regulate many different steps of gene expression, and introducing a frameshift in a *Paramecium* coding sequence can trigger RNAi<sup>28</sup>. Together with the genetic link that has been uncovered between the RNAi and NMD pathways in *C. elegans*<sup>29</sup> and in *A. thaliana*<sup>30</sup>, this raises the theoretical possibility that a translation test in the cytoplasm, which will trigger NMD in many cases of intron retention, couples mRNA degradation with the formation of RNA signal molecules that can feed back to the nucleus to modulate the splicing of homologous pre-mRNAs. Whether NMD simply allows the selection of correctly spliced transcripts or whether it has some more active function in

products from spliced (S) and unspliced (U) mRNAs from wild-type cells (no silencing), or after silencing of *UPF1A* and *UPF1B* genes or of the unrelated *ND7* gene, were resolved on agarose gels (negative of ethidium bromide stain). RT–, control reactions without reverse transcriptase. The fraction of unspliced mRNAs was quantified with ethidium bromide signals.

the choice of splice sites, our results suggest that this ancient mechanism may have evolved together with spliceosomal introns and the need to control splicing patterns.

## METHODS SUMMARY

**Bioinformatic analyses.** Intron sets from all species were confirmed by the alignment of cDNAs or ESTs, except for *S. pombe*. Only GT/AG or GC/AG introns shorter than 5,000 nt were considered. A minor fraction of gene models containing in-frame stop codons in the coding sequences were excluded from all data sets. When cDNA sequences revealed alternative splicing, each intron form was counted only once.

***Paramecium* strain, cultivation, and RNAi treatment.** The entirely homozygous strain 51 was grown in a wheatgrass-powder infusion medium bacterized with *Klebsiella pneumoniae* the day before use, and supplemented with 0.8 mg l<sup>-1</sup>  $\beta$ -sitosterol. RNAi treatment was conducted by the feeding technique: cells were cultured for seven days on the same medium containing ampicillin at 0.1 mg ml<sup>-1</sup> and bacterized with the HT115 *E. coli* strain, which produces double-stranded RNA from any sequence cloned into plasmid L4440 after induction with isopropyl  $\beta$ -D-thiogalactoside (IPTG). Sequences used for silencing of the *UPF1A*, *UPF1B*, *UPF2*, *ND7* and *ICL7a* genes were segments 1,885–2,289, 1,887–2,285, 1,143–1,546, 870–1,266 and 1–580 of the genes (from the ATG), respectively. These genes can be accessed with ParameciumDB (<http://paramecium.cgm.cnrs-gif.fr/>) under accession numbers GSPATG00034062001, GSPATG00037251001, GSPATG00017015001, GSPATG00002403001 and GSPATG00021610001, respectively.

**Northern blot analyses and RT–PCR quantification of unspliced mRNAs.** Total RNA was extracted from cells grown on *K. pneumoniae* or the relevant feeding *E. coli* strains with the use of the TRIzol (Invitrogen) procedure, modified by the addition of glass beads. Northern blots, reverse transcription and PCR were performed with standard procedures. In the RT–PCR assay, the small length difference between the spliced and unspliced versions is unlikely to have biased the PCR reaction. Any possible bias would be the same in all samples, so that it would not affect the ratio of unspliced fractions between the *UPF1A*/*UPF1B* and *ND7* silencing conditions.

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

Received 30 September; 21 November 2007.

- Roy, S. W. & Gilbert, W. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Rev. Genet.* 7, 211–221 (2006).

2. Berget, S. M. Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270**, 2411–2414 (1995).
3. Lim, L. P. & Burge, C. B. A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl Acad. Sci. USA* **98**, 11193–11198 (2001).
4. Ast, G. How did alternative splicing evolve? *Nature Rev. Genet.* **5**, 773–782 (2004).
5. Conti, E. & Izaurralde, E. Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. *Curr. Opin. Cell Biol.* **17**, 316–325 (2005).
6. Maquat, L. E. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nature Rev. Mol. Cell Biol.* **5**, 89–99 (2004).
7. Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA* **100**, 189–192 (2003).
8. Pan, Q. *et al.* Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev.* **20**, 153–158 (2006).
9. Russell, C. B., Fraga, D. & Hinrichsen, R. D. Extremely short 20–33 nucleotide introns are the standard length in *Paramecium tetraurelia*. *Nucleic Acids Res.* **22**, 1221–1225 (1994).
10. Aury, J. M. *et al.* Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444**, 171–178 (2006).
11. Romfo, C. M., Alvarez, C. J., van Heeckeren, W. J., Webb, C. J. & Wise, J. A. Evidence for splice site pairing via intron definition in *Schizosaccharomyces pombe*. *Mol. Cell Biol.* **20**, 7955–7970 (2000).
12. Ishigaki, Y., Li, X., Serin, G. & Maquat, L. E. Evidence for a pioneer round of mRNA translation: mRNAs subject to nonsense-mediated decay in mammalian cells are bound by CBP80 and CBP20. *Cell* **106**, 607–617 (2001).
13. Galvani, A. & Sperling, L. RNA interference by feeding in *Paramecium*. *Trends Genet.* **18**, 11–12 (2002).
14. Lynch, M. The origins of eukaryotic gene structure. *Mol. Biol. Evol.* **23**, 450–468 (2006).
15. Mohn, F., Buhler, M. & Muhlemann, O. Nonsense-associated alternative splicing of T-cell receptor beta genes: no evidence for frame dependence. *RNA* **11**, 147–156 (2005).
16. Wang, J., Chang, Y. F., Hamilton, J. I. & Wilkinson, M. F. Nonsense-associated altered splicing: a frame-dependent response distinct from nonsense-mediated decay. *Mol. Cell* **10**, 951–957 (2002).
17. Wang, J., Hamilton, J. I., Carter, M. S., Li, S. & Wilkinson, M. F. Alternatively spliced TCR mRNA induced by disruption of reading frame. *Science* **297**, 108–110 (2002).
18. Miriam, E., Sperling, R., Sperling, J. & Motro, U. Regulation of splicing: the importance of being translatable. *RNA* **10**, 1–4 (2004).
19. Wachtel, C., Li, B., Sperling, J. & Sperling, R. Stop codon-mediated suppression of splicing is a novel nuclear scanning mechanism not affected by elements of protein synthesis and NMD. *RNA* **10**, 1740–1750 (2004).
20. Maquat, L. E. NASTy effects on fibrillin pre-mRNA splicing: another case of ESE does it, but proposals for translation-dependent splice site choice live on. *Genes Dev.* **16**, 1743–1753 (2002).
21. Wilkinson, M. F. & Shyu, A. B. RNA surveillance by nuclear scanning? *Nature Cell Biol.* **4**, E144–E147 (2002).
22. Buhler, M., Wilkinson, M. F. & Muhlemann, O. Intranuclear degradation of nonsense codon-containing mRNA. *EMBO Rep.* **3**, 646–651 (2002).
23. Iborra, F. J., Escargueil, A. E., Kwek, K. Y., Akoulitchiev, A. & Cook, P. R. Molecular cross-talk between the transcription, translation, and nonsense-mediated decay machineries. *J. Cell Sci.* **117**, 899–906 (2004).
24. Brogna, S., Sato, T. A. & Rosbash, M. Ribosome components are associated with sites of transcription. *Mol. Cell* **10**, 93–104 (2002).
25. Dahlberg, J. E. & Lund, E. Does protein synthesis occur in the nucleus? *Curr. Opin. Cell Biol.* **16**, 335–338 (2004).
26. Iborra, F. J., Jackson, D. A. & Cook, P. R. The case for nuclear translation. *J. Cell Sci.* **117**, 5713–5720 (2004).
27. Nathanson, L., Xia, T. & Deutscher, M. P. Nuclear protein synthesis: a re-evaluation. *RNA* **9**, 9–13 (2003).
28. Garnier, O., Serrano, V., Duharcourt, S. & Meyer, E. RNA-mediated programming of developmental genome rearrangements in *Paramecium tetraurelia*. *Mol. Cell Biol.* **24**, 7370–7379 (2004).
29. Domeier, M. E. *et al.* A link between RNA interference and nonsense-mediated decay in *Caenorhabditis elegans*. *Science* **289**, 1928–1931 (2000).
30. Arciga-Reyes, L., Wootton, L., Kieffer, M. & Davies, B. UPF1 is required for nonsense-mediated mRNA decay (NMD) and RNAi in *Arabidopsis*. *Plant J.* **47**, 480–489 (2006).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank V. Wood, P. Mooney and A. Tivey for providing gff files for *S. pombe* data, and D. Gogendeau and J. Beisson for the gift of the *ICL7a* feeding plasmid. This work was funded by the CNRS and by the Agence Nationale de la Recherche. K.B. was supported by a postdoctoral contract from the CNRS. Experimental work was supported by grants from the Ministère de la Recherche and the Association pour la Recherche sur le Cancer.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence and requests for materials should be addressed to E.M. ([emeyer@biologie.ens.fr](mailto:emeyer@biologie.ens.fr)).

## METHODS

**P. tetraurelia data set.** The EST-confirmed set was constituted by selecting from the published annotation<sup>10</sup> the 6,513 gene models for which each of the predicted introns was confirmed by the alignment of at least one EST. EST alignment was as described in ref. 10. Then, 376 gene models were excluded because ESTs revealed introns that were not predicted by the annotation, a possible source of error in the identification of the reading frame. Exclusion of the 376 problematic gene models did not significantly alter the size distribution, because  $3n$  introns are also under-represented in these genes (21.6% of the total, including introns that had been overlooked in the annotation, in contrast with 40.3% and 38.0% for  $3n + 1$  and  $3n + 2$  introns, respectively).

**H. sapiens data set.** The genome sequence and the Known Genes set<sup>31</sup> were downloaded from the UCSC (University of California Santa Cruz) genome browser (<http://genome.ucsc.edu>)<sup>32,33</sup>. The version of the genome sequence is NCBI hg17 (May 2004). The Known Genes set is based on data from UniProt (SWISS-PROT and TrEMBL) and mRNA data from the NCBI (National Center for Biotechnology Information) reference sequences collection (RefSeq)<sup>34</sup> and GenBank. Observations were confirmed with genes manually annotated from the Vega consortium on chromosomes 6, 7, 9, 10, 13, 14, 20, 22 and X (<http://vega.sanger.ac.uk>)<sup>35</sup> (data not shown).

**A. thaliana data set.** We used the genome sequence and annotation from the TIGR5 release. The sequence (filename ATH1\_chr\_all.5con.gz) was downloaded from [ftp://ftp.tigr.org/pub/data/a\\_thaliana/ath1/SEQUENCES](ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES). The gene annotations were retrieved with the biomart web server ([www.biomart.org](http://www.biomart.org)) at [www.gemene.org](http://www.gemene.org). Predicted introns were confirmed by alignment with cDNA sequences. In all, 98,490 mRNA sequences from NCBI (excluding ESTs) were aligned with the genome sequence with blat<sup>36</sup>. For each mRNA sequence we selected the best genomic locus on the basis of blat scores. Alignments were filtered by selecting those with a score greater than 80% of the highest and greater than 50. Each mRNA was then realigned with the corresponding genomic region with est2genome<sup>37</sup>. The cDNA-confirmed set contained 21,233 gene models from the TIGR5 annotation, containing 110,629 introns, all confirmed by the alignment of at least one mRNA (same splice sites). A total of 386 gene models were excluded because cDNAs revealed introns that were not annotated (395 introns). The final set contained 20,847 genes and 108,783 introns.

**D. melanogaster data set.** We used release 4 (April 2004) of the genome assembly distributed by the UCSC genome browser, and the FlyBase annotation (release 4.2, September 2005). We established that 99.7% of intron annotations are validated by the alignment of at least one mRNA from RefSeq<sup>34</sup>.

**C. elegans data set.** We used the March 2004 genome assembly distributed by the UCSC genome browser, which is based on sequence version WS120 deposited into WormBase ([www.wormbase.org](http://www.wormbase.org)) as of 1 March 2004, and the WormBase gene annotation. The WormBase genes correspond to gene predictions from the WormBase WS120 files downloaded from the Sanger Institute FTP site ([ftp://ftp.sanger.ac.uk/pub/wormbase/FROZEN\\_RELEASES/WS120/CHROMOSOMES/](ftp://ftp.sanger.ac.uk/pub/wormbase/FROZEN_RELEASES/WS120/CHROMOSOMES/)).

**S. pombe data set.** Genome assembly and gene annotations were obtained from the Sanger Centre ([http://www.sanger.ac.uk/Projects/S\\_pombe/](http://www.sanger.ac.uk/Projects/S_pombe/)). The set of

EST- or cDNA-confirmed introns was built by extracting intron annotations having the tag 'confirmed'.

**General treatment of data sets.** Only GT/AG or GC/AG introns shorter than 5,000 nt were considered in all species. Gene models containing in-frame stop codons in the coding sequences were excluded from all data sets where they occurred: *H. sapiens*, 820 models; *D. melanogaster*, 42 models; *C. elegans*, 12 models; *S. pombe*, 34 models. When cDNA sequences from these species revealed alternative splicing, each intron form was counted only once.

**Reference genes.** As negative controls for the RNAi experiments, we used two genes that are not involved in NMD: ND7 and ICL7a. ND7 is involved in the control of exocytosis<sup>38</sup>. ICL7a encodes a cytoskeletal protein<sup>39</sup>.

**RT-PCR quantification of unspliced mRNAs.** Total RNA was extracted from cells grown on *K. pneumoniae* or the relevant feeding *Escherichia coli* strains with the TRIzol (Invitrogen) procedure, modified by the addition of glass beads. mRNA reverse transcription was performed with the SuperScript II kit (Invitrogen) and the anchor-oligo(dT) primer 5'-GCTCGGACCGTGGCTA-GCATTAGTGAGTTTTTTTTTTTTTTTTT-3'. After alkaline lysis of RNA and removal of the oligo(dT) primer with Microcon YM-100 centrifugal devices (Millipore), short segments containing the introns of interest were amplified by PCR with the primers listed in Supplementary Table 3, either directly from the reverse transcriptase products or, if necessary, after a first amplification with a primer corresponding to the anchor sequence and the upstream primer. For those samples in which both bands were clearly visible, the fraction of unspliced mRNAs was calculated by quantification of the ethidium bromide signal from each of the two bands, using unsaturated exposures of the agarose gels shown in Fig. 4 and the TINA software. Quantification of RT-PCR products by extension of <sup>32</sup>P-labelled primers (Supplementary Fig. 6) was performed with Sequencing Grade Taq DNA polymerase (Promega). Radioactive signals were quantified with the ImageGauge software.

31. Hsu, F. *et al.* The UCSC Known Genes. *Bioinformatics* **22**, 1036–1046 (2006).
32. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).
33. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
34. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).
35. Ashurst, J. L. *et al.* The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res.* **33**, D459–D465 (2005).
36. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
37. Mott, R. EST\_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13**, 477–478 (1997).
38. Skouri, F. & Cohen, J. Genetic approach to regulated exocytosis using functional complementation in *Paramecium*: identification of the ND7 gene required for membrane fusion. *Mol. Biol. Cell* **8**, 1063–1071 (1997).
39. Gogendeau, D. *et al.* Functional diversification of centrioles and cell morphological complexity. *J. Cell Sci.* **121**, 65–74 (2007).

# Structural basis of microtubule severing by the hereditary spastic paraplegia protein spastin

Antonina Roll-Mecak<sup>1</sup> & Ronald D. Vale<sup>1</sup>

Spastin, the most common locus for mutations in hereditary spastic paraplegias<sup>1</sup>, and katanin are related microtubule-severing AAA ATPases<sup>2–6</sup> involved in constructing neuronal<sup>7–10</sup> and non-centrosomal<sup>7,11</sup> microtubule arrays and in segregating chromosomes<sup>12,13</sup>. The mechanism by which spastin and katanin break and destabilize microtubules is unknown, in part owing to the lack of structural information on these enzymes. Here we report the X-ray crystal structure of the *Drosophila* spastin AAA domain and provide a model for the active spastin hexamer generated using small-angle X-ray scattering combined with atomic docking. The spastin hexamer forms a ring with a prominent central pore and six radiating arms that may dock onto the microtubule. Helices unique to the microtubule-severing AAA ATPases surround the entrances to the pore on either side of the ring, and three highly conserved loops line the pore lumen. Mutagenesis reveals essential roles for these structural elements in the severing reaction. Peptide and antibody inhibition experiments further show that spastin may dismantle microtubules by recognizing specific features in the carboxy-terminal tail of tubulin. Collectively, our data support a model in which spastin pulls the C terminus of tubulin through its central pore, generating a mechanical force that destabilizes tubulin–tubulin interactions within the microtubule lattice. Our work also provides insights into the structural defects in spastin that arise from mutations identified in hereditary spastic paraplegia patients.

*Drosophila* spastin is composed of an amino-terminal domain, a microtubule-interacting and -trafficking (MIT) domain that alone binds weakly to microtubules<sup>3</sup>, a poorly conserved linker element, and a carboxy-terminal AAA ATPase domain (Fig. 1a). The N-terminal region is not required for severing, because a MIT–AAA construct lacking this region robustly severs microtubules (Fig. 1a, b)<sup>3–5</sup>, has an ATPase rate similar to the full-length protein<sup>4</sup> and displays tight microtubule binding (Fig. 1b). The N-terminal region also may not be expressed in all spastin isoforms (see Supplementary Information). A segment of the poorly conserved linker (residues 390–442) is also not essential for robust microtubule-severing *in vivo*; however, truncation of the linker to <40 residues abolishes severing but not microtubule binding (Supplementary Fig. 1). The AAA construct has weak severing, ATPase and microtubule-binding activities compared with a longer construct containing the AAA and MIT domains (Fig. 1b and Supplementary Fig. 1). These results differ from a recent study<sup>14</sup> that concluded that the MIT domain is not involved in microtubule-severing.

We solved the X-ray structure of the nucleotide-free, monomeric AAA domain of *Drosophila* spastin (residues 464–758) at 2.7 Å resolution ( $R_{\text{free}} = 28.7\%$ ; Supplementary Information). Similar to other AAA proteins, the enzymatic core of spastin contains a central  $\alpha/\beta$  nucleotide-binding domain (NBD) and a smaller four-helix bundle

domain (HBD). A marked feature of the spastin structure is its open nucleotide pocket, which explains the absence of a bound nucleotide, despite the presence of 0.5 mM adenosine 5'-O-(3-thiotriphosphate) (ATP $\gamma$ S) in the crystallization solution. Comparison of our nucleotide-free spastin structure with the ATP-bound structure of *N*-ethylmaleimide-sensitive fusion protein (NSF) (an AAA protein involved in membrane fusion<sup>15</sup>) reveals that an extended loop involved in nucleotide contact and protomer–protomer interactions in NSF (Supplementary Fig. 2) is pulled away from the nucleotide pocket in spastin by the packing of the linchpin Trp 482 in a conserved hydrophobic pocket (Fig. 1g). The pocket for Trp 482 is sub-optimal, compatible with movement of the tryptophan and rearrangement of the flap on ATP-induced hexamerization or/and substrate binding.

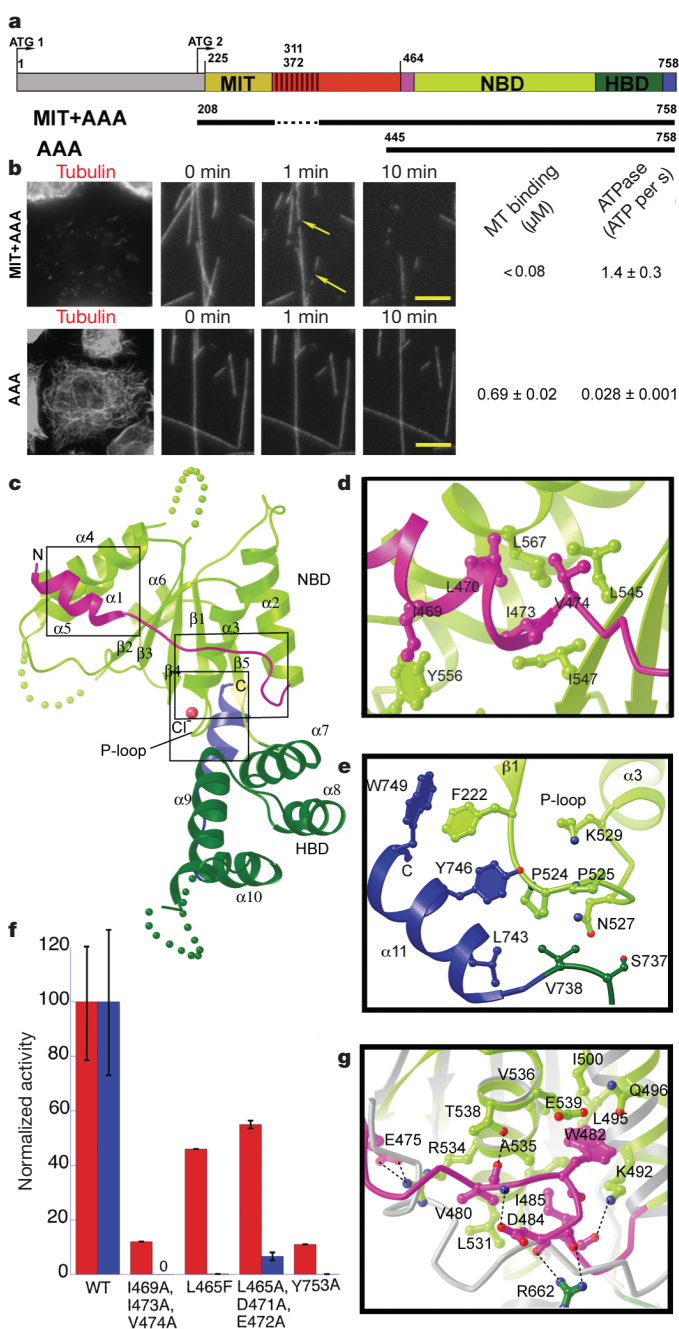
Uniquely among known AAA structures, spastin has two helices (N-terminal  $\alpha 1$  and C-terminal  $\alpha 11$ ) that embrace the NBD (Fig. 1c and Supplementary Fig. 3). The amphipathic N-terminal helix is anchored to the body of the NBD by interdigitating hydrophobic residues (Leu 470/Ile 473/Val 474; Fig. 1d). Mutation of these residues to alanine reduced ATPase activity by ~90% and abolished microtubule-severing, while preserving microtubule binding (Fig. 1f and Supplementary Fig. 4). Mutation of invariant Leu 567 located at the helix  $\alpha 1$ –NBD interface causes hereditary spastic paraplegias (HSP)<sup>16</sup>. Solvent-exposed residues on the N-terminal helix also have important roles; L465F and the triple mutant L465A/D471A/E472A markedly decreased microtubule-severing (Fig. 1f) without significantly affecting the ATPase. The C-terminal helix, which is also present in the closely related enzyme VPS4 (vacuolar sorting protein 4)<sup>17</sup>, and part of the preceding conserved linker wrap around the phosphate-binding loop (P loop) of the NBD (Fig. 1e). Mutation of the highly conserved Tyr 753 at the end of the C-terminal helix to alanine effectively inactivated the enzyme (Fig. 1f), whereas a Y753F mutation still showed severing activity *in vivo* (Supplementary Fig. 4). Thus, our structural and mutational analyses indicate that helices  $\alpha 1$  and  $\alpha 11$  of spastin have important roles in allosteric control of the ATP-binding site and possibly in substrate binding (discussed below).

We next obtained structural information on a hexameric spastin construct using small-angle X-ray scattering (SAXS). A *Caenorhabditis elegans* MIT–AAA construct was used because it is monodisperse at the concentrations (>5 mg ml<sup>-1</sup>) required for collecting high-quality SAXS data. Compared to *Drosophila* spastin, *C. elegans* spastin lacks an N-terminal domain and has a shorter linker between the MIT and AAA domains; nonetheless, it displays microtubule-severing activity<sup>18</sup> (Supplementary Fig. 5a). We first examined the oligomeric state of *C. elegans* spastin in its nucleotide-free (apo) and ATP-bound states by static multi-angle light scattering. To create a stable, noncycling ATP-bound state, we prepared a

<sup>1</sup>Howard Hughes Medical Institute and Department of Cellular and Molecular Pharmacology, University of California, San Francisco, 600 16th Street, San Francisco, California 94158, USA.

well-described AAA mutation that blocks nucleotide hydrolysis (E278Q; E583Q in *Drosophila* spastin). Static light scattering revealed that the apoenzyme exists in equilibrium between a monomeric and a weak dimeric state, whereas ATP-bound spastin is a hexamer, a quaternary structure adopted by many AAA proteins<sup>19</sup> (Supplementary Fig. 5). Unlike many AAA ATPases, but similar to katanin<sup>20</sup>, spastin exists mostly as a monomer at submicromolar concentrations, even in the presence of ATP (data not shown).

SAXS data were used to generate low-resolution *ab initio*<sup>21</sup> models of three-dimensional arrangements of scattering centres that provide the shape of the molecular envelope of the hexamer (Methods; Fig. 2 and Supplementary Fig. 6). The models from seven independent *ab initio* simulations were aligned, averaged and filtered on the basis of occupancy to obtain a most probable model. The close agreement between the total volume enclosed by the superposition of the individual runs (the composite structure) and the most probable density map (the filtered structure) indicates the robustness of the *ab initio* reconstructions. The filtered structure shows a central ring with a



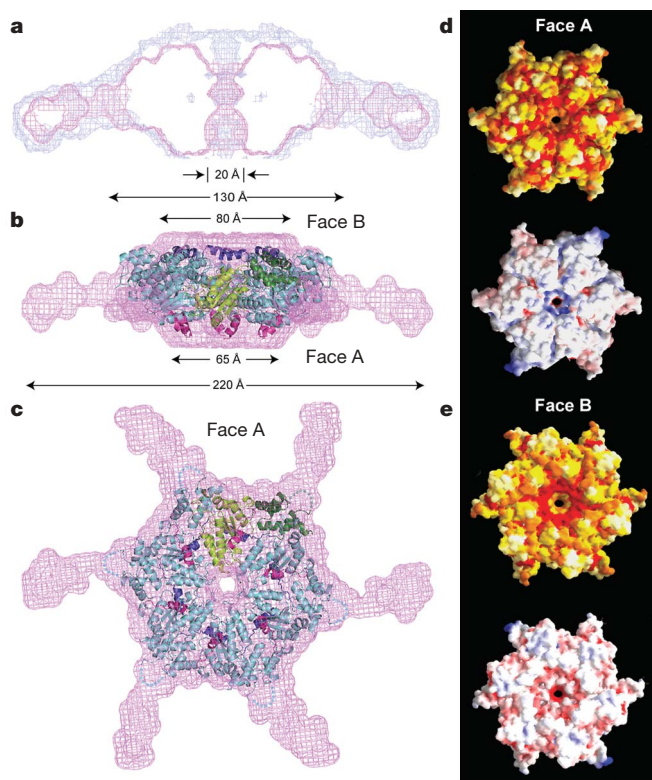
double trapezoid cross-section ( $130 \text{ \AA} \times 65 \text{ \AA}$ ), a  $\sim 20\text{-\AA}$ -diameter central pore, and slender arms radiating  $\sim 50 \text{ \AA}$  outward and extending towards one face of the ring (Fig. 2). The clear reconstruction of the arms also indicates that the linker, although unlikely to be rigid, adopts some defined structure and is not completely disordered. Shortening the linker to <40 residues disables microtubule-severing (but not microtubule-binding, Supplementary Fig. 1), suggesting some length and/or sequence requirement for this region. The asymmetric position of the arms defines a polarity to the overall structure (two faces, herein termed face A and B). We generated an atomic model for the AAA hexameric core of spastin by superimposing our nucleotide-free spastin monomer X-ray structure onto the crystal structure of the NSF hexamer. This model was docked into the SAXS reconstruction with the N- and C-terminal helices on faces A and B, respectively (Figs 2b,c; for details about the fit, see Supplementary Fig. 6 and Supplementary Information).

Several AAA proteins (for example, the bacterial proteins ClpX, ClpA and ClpB) remodel their substrates by threading the end of the polypeptide chain through a central pore in their rings<sup>19,22,23</sup>. The microtubule-severing activities of spastin and katanin depend on the  $\sim 20$ -residue disordered and negatively charged C-terminal tails of tubulin<sup>3,6</sup>, suggesting an analogous mechanism for spastin and katanin. In support of this model, we found that a 23-mer peptide corresponding to the C-terminal tail of  $\beta$ -tubulin inhibited microtubule-severing by  $\sim 70\%$  at  $0.5 \text{ mM}$ , whereas a randomized (scrambled) peptide of identical amino acid composition or an  $\alpha$ -tubulin peptide that contains the C-terminal tyrosine ( $\alpha$ -Tyr peptide) did not show detectable effects (Fig. 3a). The large concentration of peptide needed to observe inhibition is not surprising given the high local concentration of tubulin tails encountered by microtubule-bound spastin. Involvement of the  $\beta$ -tubulin tail is consistent with genetic data showing that a charge-reversal mutation in this region suppresses the lethality of ectopic katanin activity<sup>24</sup>. We also found that an antibody that recognizes exposed glutamate residues on the C-terminal tails of tubulin (detyrosinated  $\alpha$ -tubulin with a final C-terminal glutamate as well as  $\beta$ -tubulin and polyglutamylated tubulin) completely inhibited spastin-mediated severing. In contrast, a 'Tyr' antibody that recognizes  $\alpha$ -tubulin with a

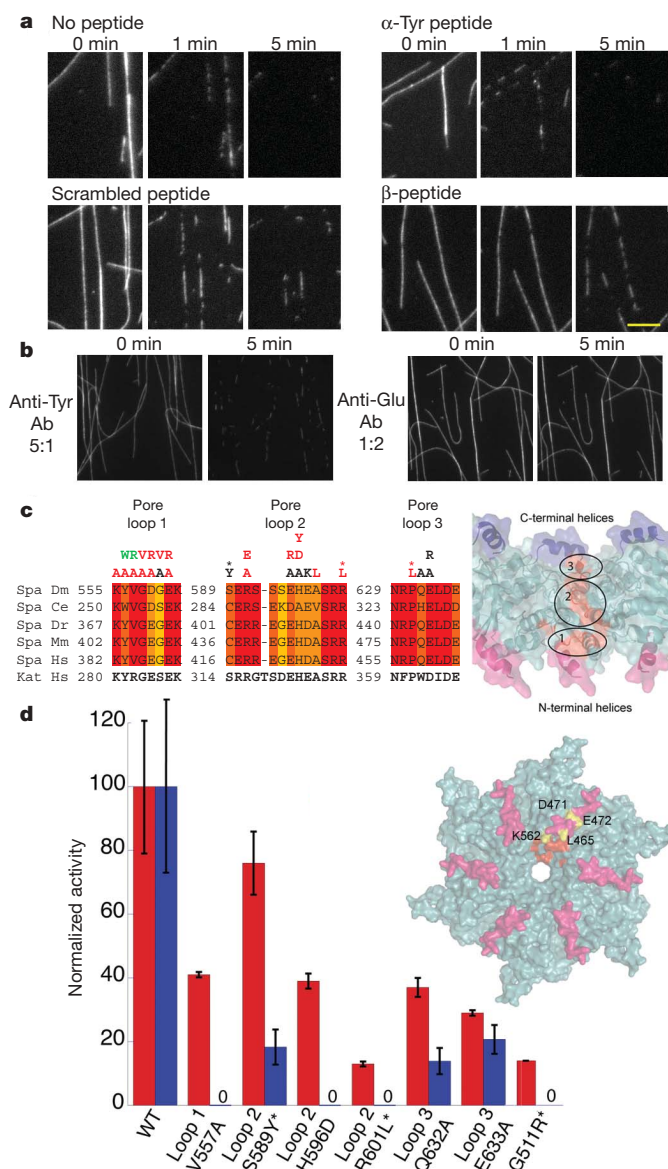
**Figure 1 | X-ray structure of the nucleotide-free AAA domain of spastin.** **a**, Domain structure of *Drosophila* spastin: grey, N-terminal domain; red, linker (exon 4, absent in the shorter isoform of spastin used in this study, is hatched); and the AAA domain (coloured according to the X-ray structure). NBD, nucleotide-binding domain; HBD, four-helix bundle domain. Two potential start codons (ATG) are shown (see Supplementary Methods for discussion). The N-terminal boundary of the AAA domain is based on our X-ray structure and differs from that of ref. 14. A segment of the structurally important N-terminal helix of the AAA domain is within what the authors of ref. 14 define as a microtubule-binding domain. The MIT + AAA and AAA constructs are shown schematically below. **b**, Left, MIT + AAA disassembles the microtubule network when transfected in *Drosophila* S2 cells and when added to microtubules *in vitro*, but AAA has no detectable activity at the same concentration ( $0.15 \mu\text{M}$ ). (Weak severing is observed at higher concentrations, Supplementary Fig. 1.) Arrows indicate breaks in microtubules. Scale bar,  $5 \mu\text{m}$ . Right, microtubule (MT)-binding and ATPase activities of MIT + AAA and AAA. Microtubule-binding affinity was determined for the Walker B E583Q mutant, which is a stable hexamer and is inactive in severing. **c**, Ribbon representation of the spastin AAA domain crystal structure. N-terminal helix/loop, magenta; NBD, light green; HBD, dark green; C-terminal helix, blue. The pink sphere depicts a chloride ion. **d**, Conserved hydrophobic interactions between the N-terminal helix and the main body of the NBD. **e**, Conserved interactions between the C-terminal helix and the P loop. **f**, ATPase (red) and microtubule-severing (blue) rates of N- and C-terminal helix mutants. Error bars represent standard errors of the mean (see Methods). WT, wild type. **g**, Detail of the superposition of spastin and ATP-bound NSF structures<sup>15</sup>, showing contacts that keep the N-terminal flap of monomeric spastin (magenta) in an open conformation, unable to stabilize the nucleotide or interact with the neighbouring protomer. Spastin is colour-coded as in panel c. NSF is in grey. Dashed lines, hydrogen bonds.

C-terminal tyrosine<sup>25</sup> (~50% of brain tubulin<sup>26,27</sup>) did not inhibit severing, even though the antibody binds to microtubules (Fig. 3b and Supplementary Fig. 7b). Although we did not detect a robust inhibitory effect of a detyrosinated  $\alpha$ -tubulin peptide, an antibody that recognizes the tail of Glu- $\alpha$ -tubulin<sup>27</sup> partially inhibited severing (Supplementary Fig. 7c). Collectively, these *in vitro* data support a model in which spastin interacts with the acidic tubulin C-terminal peptide during the severing reaction and may recognize specific features of the C-terminal peptide.

To explore this model further, we examined the roles of three solvent-exposed loops within the pore that are highly conserved among spastins and katanins (Fig. 3c). Mutations in pore loop 1 of *Drosophila* spastin, which has been shown to be important for the substrate-remodelling activity of several other AAA proteins<sup>22,23,28</sup>, abolished severing (Figs 3c, d) but preserved microtubule binding (Supplementary Table 2 and Supplementary Fig. 8). After submission of this work, similar results were obtained in ref. 14. Mutations of solvent-exposed residues in pore loops 2 and 3 also completely inhibited or severely crippled the enzyme (Figs 3c, d). However, the disease mutant S589Y retains some activity, suggesting neurons



**Figure 2 | Model of active, hexameric spastin from light and small-angle X-ray scattering.** **a**, *Ab initio* SAXS reconstructions<sup>21</sup> of *C. elegans* spastin (MIT + AAA; residues 15–452, ATP-hydrolysis-deficient E278Q mutant). Shown is a cross-section through the filtered (magenta) and composite (blue) SAXS envelopes. The composite structure consists of the aligned, superimposed and summed models from seven independent simulations, whereas the filtered model corresponds to the most probable density map. **b**, **c**, Fit of a spastin hexameric model into the SAXS reconstruction (equatorial (**b**) and axial (**c**) views). In the absence of an atomic model of the MIT + linker, its precise location within the envelope is uncertain. Maximal diameter is given at various heights of the structure. **c** shows face A of the hexamer. For details, see Methods. Colour-coding for spastin is as in Fig. 1c. **d**, **e**, Surface properties of face A (**d**) and face B (**e**). Top image of **d** and **e**, solvent-accessible surface of the spastin hexamer model, colour-coded for amino acid similarity as in Supplementary Fig. 3 (white, 40% identity, to dark red, 100% identity, among spastin and katanins). Bottom, solvent-accessible surface of the spastin hexamer model, colour-coded for electrostatic potential (red, negative; blue, positive, ranging from –12 kT to 12 kT).



**Figure 3 | Role of the tubulin C terminus and the spastin pore in microtubule-severing.** **a**, Effects of tubulin C-terminal peptides on microtubule-severing *in vitro*. Addition of  $\alpha$ -Tyr peptide had no detectable effect on severing rates, whereas a  $\beta$ -tubulin C-terminal peptide reduced the severing rate. A scrambled peptide had no detectable effect (see Methods). Scale bar, 5  $\mu$ m. **b**, Antibodies (Ab) recognizing Glu- $\alpha$ -tubulin,  $\beta$ -tubulin and polyglutamylated tubulin (anti-‘Glu’) inhibit spastin-mediated severing completely at 1:2 antibody:tubulin molar ratio (the same level of protection was seen even 30 min after spastin addition), whereas antibodies recognizing Tyr- $\alpha$ -tubulin (anti-Tyr) did not protect against severing, even at a 5:1 antibody:tubulin molar ratio (antibody binding to these microtubules demonstrated in Supplementary Fig. 7b). **c**, Left, conservation of the three pore loops. Loop 1 residues are conserved in all AAA ATPases; loops 2 and 3 are specific to the spastin subfamily. Effects on microtubule-severing of mutations in pore loop residues are shown on top of the alignment: red, inactive; black, severely crippled; green, active. Asterisks denote disease mutations. The effects of mutations generally decrease in severity from the pore entrance to the exit, with loop 1 being the least permissive to substitutions. Right, positions of pore loops (labelled 1, 2 and 3) in the spastin hexamer, in a cross-sectional view of the pore. Side-chains for residues K555, Y556 and D559 as well as residues 592–596 are not visible in the electron density maps and are presumed to be disordered. **d**, Left, ATPase (red) and microtubule-severing (blue) rates for selected mutants. Error bars represent standard errors of the mean (see Methods). Right, molecular surface of the hexameric spastin model showing in yellow the location of residues on face A that impair severing. Loop residues that impair severing are shown in red.

are susceptible to disease with partial spastin activity. Mutations of surface residues leading to the pore (Fig. 3d) also markedly affected the activity of spastin (for example, L465F and L465A/D471A/E472A in Fig. 1f, and K562A and K562R in Supplementary Fig. 4).

In conclusion, the combination of X-ray crystallography, SAXS *ab initio* reconstructions and structure-guided mutagenesis provides the first structural information on microtubule-severing proteins and allows us to propose a molecular model for spastin-mediated severing (Fig. 4a). Owing to their similar domain organization and high sequence similarity, this model probably pertains to katanin as well. We propose that face A of the spastin AAA ring docks onto the microtubule, placing the positively charged N-terminal pore entrance in contact with the negatively charged C terminus of tubulin. The translocation from face A to face B would correspond to the direction of substrate translocation proposed for the distantly related AAA ATPases ClpX, ClpA and ClpB<sup>22,28,29</sup>. The linker and MIT domains extending from the ring would make additional contacts with the microtubule, thus increasing microtubule avidity and potentially stabilizing the hexamer on the microtubule<sup>20</sup>. On the

basis of our affinity measurements, only a subset of the six arms is likely to make strong binding interactions with the microtubule (Fig. 4a).

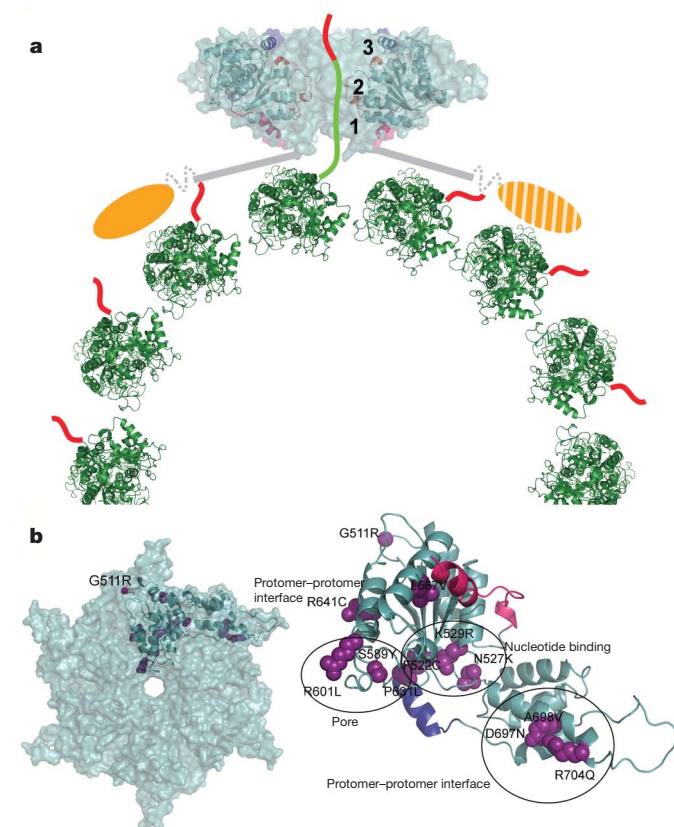
We propose that the tubulin polypeptide is threaded through the pore, perhaps driven by nucleotide-driven conformational changes of the pore loops. However, spastin may not need to completely translocate the tubulin polypeptide substrate, but instead just grip the C-terminal tubulin tail and exert mechanical 'tugs' that might partially unfold tubulin or locally destabilize protomer-protomer interactions, leading to catastrophic breakdown of the microtubule lattice. It also remains possible that the MIT domains could participate in this nucleotide-driven process by binding and 'feeding' the C-terminal tails to the pore. Further biophysical characterization will be needed to decipher the structural details of substrate recognition and mechanical force production. Our data also suggest that spastin may selectively recognize post-translationally modified tubulins ('Glu' tubulins) that are part of stable microtubules. Consistent with this idea, loss of spastin in *Drosophila* results in the accumulation of stabilized polyglutamylated tubulin in neurons<sup>8</sup> and spastin knock-out mice show axonal swellings enriched in dephosphorylated, stable microtubules<sup>30</sup>. Our structure also provides the first glimpse into how spastin disease mutations contribute to spastin dysfunction and disease, most of which we suggest are involved in destabilizing protomer-protomer interactions, microtubule- or ATP-binding (Fig. 4b and Supplementary Fig. 4); in such cases, spastin-linked HSP is probably caused by haploinsufficiency and not a dominant negative effect. Further elucidation of the mechanistic details of how spastin interacts with particular tubulin isoforms and post-translational modifications and leads to microtubule destabilization may provide insight into the origin of spastin paraplegias and potential treatments for this disease.

## METHODS SUMMARY

Crystallographic statistics can be found in Supplementary Table 1.

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

Received 23 March; accepted 16 November 2007.



**Figure 4 | Proposed mechanism of severing by spastin and effects of disease mutations.** **a**, Proposed mechanism for microtubule-severing by spastin. The spastin AAA core is shown in cyan with pore loops 1, 2 and 3 highlighted in red and numbered in the figure. The MIT domains are shown as gold ovals. The valency of the interaction of the MIT domains with the microtubule is unknown. On the basis of affinity measurements, it is likely that not all MIT domains are engaged with the microtubule (the potentially unengaged MIT domain is shown hatched). The tubulin heterodimers forming the microtubule are shown in green as a ribbon representation, whereas the C-terminal tubulin tails are shown in red cartoon representation. **b**, Left, molecular surface of spastin (face A). One protomer is shown in a ribbon representation and residues mutated in HSP patients are shown as violet spheres. Right, in addition to mapping to the pore loops (S589Y, R601L, P631L), disease mutations can interfere with ATP binding (F522C, N527K, K529R) and protomer-protomer interactions (D697N, R704Q, R641C, R601L, P631L). G511R maps to a loop on face A where it could destabilize protomer-protomer interactions and/or the microtubule-binding interface (Supplementary Fig. 4).

1. Hazan, J. *et al.* Spastin, a new AAA protein, is altered in the most frequent form of autosomal dominant spastic paraplegia. *Nature Genet.* **23**, 296–303 (1999).
2. Frickey, T. & Lupas, A. N. Phylogenetic analysis of AAA proteins. *J. Struct. Biol.* **146**, 2–10 (2004).
3. Roll-Mecak, A. & Vale, R. D. The *Drosophila* homologue of the hereditary spastic paraplegia protein, spastin, severs and disassembles microtubules. *Curr. Biol.* **15**, 650–655 (2005).
4. Evans, K. J., Gomes, E. R., Reisenweber, S. M., Gundersen, G. G. & Lauring, B. P. Linking axonal degeneration to microtubule remodeling by Spastin-mediated microtubule severing. *J. Cell Biol.* **168**, 599–606 (2005).
5. Salinas, S. *et al.* Human spastin has multiple microtubule-related functions. *J. Neurochem.* **95**, 1411–1420 (2005).
6. McNally, F. J. & Vale, R. D. Identification of katanin, an ATPase that severs and disassembles stable microtubules. *Cell* **75**, 419–429 (1993).
7. Roll-Mecak, A. & Vale, R. D. Making more microtubules by severing: a common theme of noncentrosomal microtubule arrays? *J. Cell Biol.* **175**, 849–851 (2006).
8. Trotta, N., Orso, G., Rossetto, M. G., Daga, A. & Broadie, K. The hereditary spastic paraplegia gene, spastin, regulates microtubule stability to modulate synaptic structure and function. *Curr. Biol.* **14**, 1135–1147 (2004).
9. Sherwood, N. T., Sun, Q., Xue, M., Zhang, B. & Zinn, K. *Drosophila* Spastin regulates synaptic microtubule networks and is required for normal motor function. *PLoS Biol.* **2**, e429 (2004).
10. Wood, J. D. *et al.* The microtubule-severing protein Spastin is essential for axon outgrowth in the zebrafish embryo. *Hum. Mol. Genet.* **15**, 2763–2771 (2006).
11. Burk, D. H., Liu, B., Zhong, R., Morrison, W. H. & Ye, Z. H. A katanin-like protein regulates normal cell wall biosynthesis and cell elongation. *Plant Cell* **13**, 807–827 (2001).
12. Srayko, M., Buster, D. W., Bazirgan, O. A., McNally, F. J. & Mains, P. E. MEI-1/MEI-2 katanin-like microtubule severing activity is required for *Caenorhabditis elegans* meiosis. *Genes Dev.* **14**, 1072–1084 (2000).
13. Zhang, D., Rogers, G. C., Buster, D. W. & Sharp, D. J. Three microtubule severing enzymes contribute to the "Pacman-flux" machinery that moves chromosomes. *J. Cell Biol.* **177**, 231–242 (2007).

14. White, S. R., Evans, K. J., Lary, J., Cole, J. L. & Lauring, B. Recognition of C-terminal amino acids in tubulin by pore loops in Spastin is important for microtubule severing. *J. Cell Biol.* **176**, 995–1005 (2007).
15. Yu, R. C., Hanson, P. I., Jahn, R. & Brunger, A. T. Structure of the ATP-dependent oligomerization domain of *N*-ethylmaleimide sensitive factor complexed with ATP. *Nature Struct. Biol.* **5**, 803–811 (1998).
16. Fonknechten, N. *et al.* Spectrum of SPG4 mutations in autosomal dominant spastic paraplegia. *Hum. Mol. Genet.* **9**, 637–644 (2000).
17. Scott, A. *et al.* Structural and mechanistic studies of VPS4 proteins. *EMBO J.* **24**, 3658–3669 (2005).
18. Matsushita-Ishiodori, Y., Yamanaka, K. & Ogura, T. The *C. elegans* homologue of the spastic paraplegia protein, spastin, disassembles microtubules. *Biochem. Biophys. Res. Commun.* **359**, 157–162 (2007).
19. Sauer, R. T. *et al.* Sculpting the proteome with AAA(+) proteases and disassembly machines. *Cell* **119**, 9–18 (2004).
20. Hartman, J. J. & Vale, R. D. Microtubule disassembly by ATP-dependent oligomerization of the AAA enzyme katanin. *Science* **286**, 782–785 (1999).
21. Svergun, D. I., Petoukhov, M. V. & Koch, M. H. Determination of domain structure of proteins from X-ray solution scattering. *Biophys. J.* **80**, 2946–2953 (2001).
22. Hinnerwisch, J., Fenton, W. A., Furtak, K. J., Farr, G. W. & Horwich, A. L. Loops in the central channel of ClpA chaperone mediate protein binding, unfolding, and translocation. *Cell* **121**, 1029–1041 (2005).
23. Schlieker, C. *et al.* Substrate recognition by the AAA+ chaperone ClpB. *Nature Struct. Mol. Biol.* **11**, 607–615 (2004).
24. Lu, C., Srayko, M. & Mains, P. E. The *Caenorhabditis elegans* microtubule-severing complex MEI-1/MEI-2 katanin interacts differently with two superficially redundant beta-tubulin isotypes. *Mol. Biol. Cell* **15**, 142–150 (2004).
25. Wehland, J., Willingham, M. C. & Sandoval, I. V. A rat monoclonal antibody reacting specifically with the tyrosylated form of alpha-tubulin. I. Biochemical characterization, effects on microtubule polymerization *in vitro*, and microtubule polymerization and organization *in vivo*. *J. Cell Biol.* **97**, 1467–1475 (1983).
26. Rodriguez, J. A. & Borisy, G. G. Tyrosination state of free tubulin subunits and tubulin disassembled from microtubules of rat brain tissue. *Biochem. Biophys. Res. Commun.* **89**, 893–899 (1979).
27. Gundersen, G. G., Kalnoski, M. H. & Bulinski, J. C. Distinct populations of microtubules: tyrosinated and nontyrosinated alpha tubulin are distributed differently *in vivo*. *Cell* **38**, 779–789 (1984).
28. Siddiqui, S. M., Sauer, R. T. & Baker, T. A. Role of the processing pore of the ClpX AAA+ ATPase in the recognition and engagement of specific protein substrates. *Genes Dev.* **18**, 369–374 (2004).
29. Lee, S., Choi, J. M. & Tsai, F. T. Visualizing the ATPase cycle in a protein disaggregating machine: structural basis for substrate binding by ClpB. *Mol. Cell* **25**, 261–271 (2007).
30. Tarrade, A. *et al.* A mutation of spastin is responsible for swellings and impairment of transport in a region of axon characterized by changes in microtubule composition. *Hum. Mol. Genet.* **15**, 3544–3558 (2006).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank C. Ralston for access to beamlines at the Advanced Light Source (Lawrence Berkeley National Laboratory), G. Hura for assistance during the SAXS experiments and data processing, N. Zhang for assistance with molecular biology, D. Southword for advice with the static multi-angle scattering experiments, T. Huckaba for the anti-Glu  $\alpha$ -tubulin antibody, and H. Bourne and A. Ferre-D'Amare for support and critical reading of the manuscript. R.D.V. is a Howard Hughes Medical Institute investigator. A.R.-M. has received support from the Damon Runyon Cancer Research Foundation, the NIH and the Burroughs Wellcome Fund.

**Author Information** Atomic coordinates and structure factor amplitudes have been deposited in the Protein Data Bank under the accession number 3B9P. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence and requests for materials should be addressed to R.D.V. ([vale@cmp.ucsf.edu](mailto:vale@cmp.ucsf.edu)).

## METHODS

**X-ray structure determination.** The expression, purification, crystallization and X-ray structure determination are described in Supplementary Methods.

**Multi-angle light-scattering measurements.** These experiments are described in Supplementary Methods.

**Solution X-ray scattering data acquisition, analysis and modelling.** SAXS data for the reconstructions of the *C. elegans* spastin holoenzyme bound to ATP were collected at the SIBYLS beamline at the Advanced Light Source (ALS), Lawrence Berkeley National Laboratory (details on data collection are provided in Supplementary Methods). Data were collected at concentrations of 9 mg ml<sup>-1</sup>, 5 mg ml<sup>-1</sup> and 2.5 mg ml<sup>-1</sup>. The raw scattering data were scaled, and buffers were subtracted by using software written by G. Hura (ALS SYBILS). Individual scattering curves were visually inspected before averaging to ensure radiation damage was minimal. There was no evidence for radiation-induced aggregation. Individual scattering curves collected at different concentrations were scaled and merged in PRIMUS to yield a low-noise composite curve. The radius of gyration ( $R_g$ ) was initially computed from the Guinier plot<sup>31</sup> as 58.5 Å. The pair distance distribution function  $P(r)$  was calculated using the indirect Fourier transform method of Svergun as implemented in the program package GNOM<sup>32</sup>. The value for  $D_{max}$  was determined empirically by examining the quality of the fit to the experimental data for a range of  $D_{max}$  values from 220 Å to 235 Å. A value of 230 Å was used in the *ab initio* modelling.

The program GASBOR<sup>31</sup> was used to provide a shape for the molecular envelope of the hexamer. Because the inverse scattering problem has no unique solution, eight *ab initio* reconstructions were performed. GASBOR assigns a pseudo residue to each residue in the protein. The only information that went into the shape reconstructions, other than the X-ray scattering curve, is the number of residues to be modelled and the value of  $D_{max}$ . Six-fold symmetry was imposed. Because the simulation uses a constant number of equal-density scattering elements, the regions in the structure that are flexible are assigned different positions in the individual simulations. The eight independent reconstructions have the same overall architecture. The dummy atom models resulting from the eight individual GASBOR runs were aligned, averaged and scored with a normalized structural difference (NSD) using DAMAVER<sup>33</sup>. The criterion for inclusion in averaging procedures was  $NSD < \text{mean NSD} + 2 \times \text{variation}$ . Model 6 was discarded using this criterion (Supplementary Fig. 6). The seven selected *ab initio* models agreed well, yielding  $1.442 \pm 0.137$  (NSD  $\pm$  s.d.). A most-probable model can be generated by filtering the results of the seven independent reconstructions. The level of heterogeneity among the individual GASBOR models is apparent when the averaged and filtered model is compared to the total volume enclosed by the superposition of the models from the individual simulations (the composite structure from seven independent simulations). Figure 2 shows the aligned, superimposed and summed individual GASBOR models (light blue) and the filtered model (magenta). Supplementary Fig. 6 shows thin vertical and horizontal slices through the composite and filtered models as well as four independent reconstructions (Supplementary Fig. 6g). The comparison between the solution scattering curve (black line) and

computed scattering curves for the eight models is shown in Supplementary Fig. 6c. The fit between the atomic model of the hexameric ring and the SAXS model is shown in Fig. 2 (Supplementary Fig. 6f; details in Supplementary Information).

**Microtubule binding, ATPase and severing assays.** Microtubule binding and ATPase assays are described in Supplementary Methods. For the *in vitro* microtubule-severing assays, spastin constructs (0.15  $\mu$ M) were added to microtubule-coated flow cells with 1 mM ATP, 2 mg ml<sup>-1</sup> casein, 10  $\mu$ M taxol and an oxygen-scavenging system described in ref. 34 (details in Supplementary Information). Images were recorded every 15 s using a Zeiss inverted microscope with a Hamamatsu ORCA-AG camera. The rate of microtubule-severing was determined by measuring the rate of shortening of severed microtubules (gaps of 1  $\mu$ m to 3  $\mu$ m) from their ends over time. These measurements were done in the 'burst phase' of the reaction, before there were too many severing sites that would destabilize and unravel the microtubule lattice, probably in a spastin-independent manner. Reported rates were derived from measurements at 21–80 separate severing sites, with the exception of mutant Y753A where only seven breaks were observed owing to its low activity. Similar relative activities for the constructs were obtained by measuring the total number of microtubule breaks (scored manually from time-lapse imaging) per unit length of microtubule per unit time (not shown).

For the peptide studies, MIT + AAA *Drosophila* spastin was pre-incubated for 30 s with the indicated concentration of peptide (97% pure from SynBio Inc.) and then perfused in the flow chamber.  $\alpha$ -Tyr peptide, CEVGVDSVEGEG-EEEEEEY;  $\beta$ -tubulin C-terminal peptide, CQYQDATADEQGEFEEEGEEDEA; scrambled peptide, CQETAEYQDEEQGEADAEDFG. Data were recorded and analysed as described above. For the antibody studies, microtubules were immobilized to the glass surface and the various antibodies were perfused into the chamber and incubated for 10 min, followed by three washes and then addition of the *Drosophila* MIT + AAA construct and ATP. The antibody referred to as anti-Glu is a monoclonal antibody specific for Glu- $\alpha$ -tubulin (Synaptic Systems, CI 1D5); however, it cross-reacts with  $\beta$ -tubulin and polyglutamylated tubulin, but does not recognize tyrosinated  $\alpha$ -tubulin. The same level of protection from severing was seen even 30 min after spastin addition. The antibody referred to as anti-Tyr is a monoclonal antibody specific for tyrosinated  $\alpha$ -tubulin (Synaptic Systems, CI 20C6). The polyclonal antibody specific for Glu- $\alpha$ -tubulin<sup>27</sup> was used at a 1:10 dilution from serum (Supplementary Fig. 7). The same level of protection from severing was observed even 30 min after spastin addition.

- Guinier, A. & Fournet, G. *Small Angle Scattering of X-rays* (John Wiley and Sons, New York, 1995).
- Svergun, D. I. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J. Appl. Crystallogr.* **25**, 495–503 (1992).
- Volkov, V. V. & Svergun, D. I. Uniqueness of *ab initio* shape determination in small-angle scattering. *J. Appl. Crystallogr.* **36**, 860–864 (2003).
- Yildiz, A., Tomishige, M., Vale, R. D. & Selvin, P. R. Kinesin walks hand-over-hand. *Science* **303**, 676–678 (2004).

# naturejobs

**JOBS OF  
THE WEEK**

**P**rospective postdocs are usually advised to pick a fellowship outside their home country. And the winners of *Naturejobs's* Postdoc Journal contest this year are nomadic even by postdoc standards. Such journeys come with strains. Differences in language and culture make the distance from home about more than just kilometres, and complicate a professional journey fraught with obstacles such as scientific competition, uncooperative data and learning to manage a lab.

More than 50 fellows competed to share their stories in our Postdoc Journal feature in 2008. Applicants came from around the world — including Switzerland, China, South Africa, the Philippines, Australia and Finland. All told tales of their career journey. But the four winners spun the best scientific travelogues.

Aliza le Roux is a South African primate-behaviour researcher recently arrived in the United States. She has two adjustments to make, first to working for a US university, then to fieldwork in Ethiopia, where she will be stationed as a fellow for the University of Michigan, Ann Arbor. Amanda Goh, who completed her graduate work in the United States, is now a fellow at the Biopolis in Singapore. She is adapting to Asian life, and is already fielding questions about Singapore's big science budget and reputation for social strictness. UK-born Jon Yearsley is a self-described serial postdoc, in and out of fellowships for 10 years. He's giving himself one more year as a fellow at the University of Lausanne in Switzerland, where he hopes to complete a move from theoretical cosmology to ecology and evolutionary biology. And plant geneticist Zachary Lippman of the Hebrew University of Jerusalem knows that scientific travel can be dangerous, as a few summers back his field experiments were caught in the war between Israel and Hezbollah in Lebanon.

We'd like to thank all the applicants willing to share their travails, and congratulate the four winners. We wish them all a safe journey and a satisfying arrival — even as we anticipate reading about them navigating the bumps in their roads.

**Gene Russo, acting editor of *Naturejobs***

## CONTACTS

**Acting Editor:** Gene Russo

**European Head Office, London**  
The Macmillan Building,  
4 Crinan Street, London N1 9XW, UK  
Tel: +44 (0) 20 7843 4961  
Fax: +44 (0) 20 7843 4996  
e-mail: [naturejobs@nature.com](mailto:naturejobs@nature.com)

**European Sales Manager:**  
Andy Douglas (4975)  
e-mail: [a.douglas@nature.com](mailto:a.douglas@nature.com)  
**Business Development Manager:**  
Amelie Pequignot (4974)  
e-mail: [a.pequignot@nature.com](mailto:a.pequignot@nature.com)  
**Natureevents:**

Claudia Paulsen Young  
(+44 (0) 20 7014 4015)  
e-mail: [c.paulsenyoung@nature.com](mailto:c.paulsenyoung@nature.com)  
**France/Switzerland/Belgium:**  
Muriel Lestringuez (4994)

**Southwest UK/RoW:**  
Nils Moeller (4953)  
**Scandinavia/Spain/Portugal/Italy:**  
Evelina Rubio-Hakansson (4973)  
**Northeast UK/Ireland:**  
Matthew Ward (+44 (0) 20 7014 4059)  
**North Germany/The Netherlands:**  
Reya Silao (4970)  
**South Germany/Austria:**  
Hildi Rowland (+44 (0) 20 7014 4084)

**Advertising Production Manager:**  
Stephen Russell  
To send materials use London  
address above.  
Tel: +44 (0) 20 7843 4816  
Fax: +44 (0) 20 7843 4996  
e-mail: [naturejobs@nature.com](mailto:naturejobs@nature.com)  
**Naturejobs web development:**  
Tom Hancock  
**Naturejobs online production:**  
Dennis Chuz

**US Head Office, New York**  
75 Varick Street, 9th Floor,  
New York, NY 10013-1917  
Tel: +1 800 989 7718  
Fax: +1 800 989 7103  
e-mail: [naturejobs@natureny.com](mailto:naturejobs@natureny.com)

**US Sales Manager:** Peter Bless

**Japan Head Office, Tokyo**  
Chiyoda Building, 2-37 Ichigayatamachi,  
Shinjuku-ku, Tokyo 162-0843  
Tel: +81 3 3267 8751  
Fax: +81 3 3267 8746

**Asia-Pacific Sales Manager:**  
Ayako Watanabe (+81 3 3267 8765)  
e-mail: [a.watanabe@natureasia.com](mailto:a.watanabe@natureasia.com)  
**Business Development Manager, Greater  
China/Singapore:**  
Gloria To (+852 2811 7191)  
e-mail: [g.to@natureasia.com](mailto:g.to@natureasia.com)

# MOVERS

**Leszek Borysiewicz, chief executive,  
Medical Research Council, London**



**2004-07:** Deputy rector,  
Imperial College London  
**2001-04:** Principal, Faculty  
of Medicine, Imperial College  
London  
**1991-2001:** Head,  
Department of Medicine,  
University of Wales, Cardiff

Leszek Borysiewicz's experience as doctor, researcher and an academic administrator has prepared him well to lead Britain's Medical Research Council (MRC). His research knowhow is bolstered by "extremely good management judgement", says David Delpy, chief executive of the Engineering and Physical Sciences Research Council.

After studying medicine at what is now the University of Wales, Borysiewicz went on to the Royal Postgraduate Medical School of London, where he witnessed mixed results of kidney transplants. "The kidneys were surviving, but the patients were falling foul of cytomegalovirus," Borysiewicz says. The MRC, which had links with the school, funded him on a basic-science degree that aroused his fascination about how latent viruses could morph into pathogens.

Smitten with research, he continued as a postdoc and lecturer at the school, then went as a physician to the Gambia, which sparked an interest in global health issues. After a stint at Cambridge, he returned to his hometown of Cardiff as professor of medicine at the University of Wales. There, he assembled a large team of doctors, scientists and nurses who carried out clinical trials for a therapeutic vaccine for human papillomavirus — the first in Europe. He received a knighthood for this work in 2001.

Borysiewicz was never discouraged by negative results, says Stephen Man, who worked with him at Cardiff. "He'd use them as a new avenue for investigation," Man says. "He was extremely enthusiastic."

Moving to Imperial College, London, Borysiewicz climbed the administration ladder to become deputy rector. He developed a collegial relationship with Delpy, who was vice-provost for research at University College London at the time. They regularly reviewed strategies and considered how to respond to calls from the government, says Delpy. He and Borysiewicz look forward to teaming up again as chief executives, taking on major cross-council themes such as ageing, environmental change and health care.

In October, the UK government announced a boost in health-research funding. At the MRC, this will help expand translational research, which has been a contentious issue at an institution revered for its contributions to basic science. Borysiewicz says translation can now move forward without penalizing basic science.

Working on global health, interdisciplinary research and translating basic science to benefit society: it's a heady mix. "I can't think of a more exciting job," says Borysiewicz. ■

Jill U. Adams

## SCIENTISTS & SOCIETIES

### Bound for Bangalore

Norio Kikuchi received his BSc in physics from the University of Tokyo in 2000, picked up his DPhil in theoretical physics from Oxford in 2003, and then moved on to Germany for postdoctoral research. Then he did something surprising.

Although he had several offers from the United States and Europe as well as his native Japan, he joined the Indian Institute of Science (IISc) in Bangalore as a postdoc. Since last August he has worked with a group studying soft-condensed matter. He makes just US\$625 a month, much less than he would receive elsewhere. (The cost of living is lower, though, and the IISc provides housing.)

An increasing number of young scientists are attracted to India, despite lean pay cheques. Kikuchi was drawn by the chance to work with renowned condensed-matter physicist Sriram Ramaswamy. "I like Indian culture and food, and my artist wife loves India too," says Kikuchi. "That is also an important factor."

"We still encourage Indian students to go abroad for postdoctoral training," says Jayaraman Srinivasan, head of the IISc Centre for Atmospheric and Oceanic Sciences, "and many come back. At the same time, we want researchers from other countries to come and see what our institutes can offer." Nine postdocs have joined the

IISc under a new Centenary Post-doctoral Fellowships scheme, which has received applications from other countries. Kikuchi is the first foreigner chosen. "Once the scheme gets visibility, more foreign researchers will come," says Srinivasan, whose centre already has four postdocs from France through a separate bilateral scheme. The IISc can provide 50 postdocs, says associate director Narayanaswamy Balakrishnan — more if funds become available. In two years, it will open a hostel for 100 postdocs.

Other institutions are taking the IISc's cue. This month, the Indian Council of Medical Research (ICMR), which runs more than 30 institutes, will start offering fellowships for biomedical scientists in developing countries, inviting them to work in Indian institutes and laboratories. Kanikaram Satyanarayana, deputy chief of the ICMR, says it plans to offer five fellowships a year, each lasting for one to six months with return airfare paid. One aim is better 'south to south' cooperation.

"Here I have enough time to think in a creative atmosphere, which perhaps results from Indian peoples' ways of living," says Kikuchi. "I also can focus on my work, without any unnecessary politics and paper work." ■

K. S. Jayaraman

#### POSTDOC JOURNAL

### Starting anew

I am on American soil for the first time in my life. I was offered a postdoc research position two weeks ago, quit my time-filler job, left my home in South Africa, braved a 30-hour flight and am about to embark on a venture that will take me out of my comfort zone. Starting in February, I will conduct research for the University of Michigan, studying the communication and cognition of monkeys known as geladas in the Ethiopian highlands.

I completed my PhD just five days before writing this. For the first time in my life, I do not have the protection of a degree to buffer me. While I was studying, time was flexible and success hinged on a thesis that only my examiners would ever read. Now I have a contract, and an army of peers will determine whether or not I do well. I feel utterly exposed. Will I be capable of generating truly novel hypotheses? How independent am I, really? Being a 'fellow' — not a student — sounds frightening. It also sounds exhilarating. Am I equipped to handle it?

I am tackling these questions by jumping in at the deep end. For the next two years I will be overstimulating myself in an isolated, strange place, immersing myself in a research subject that I've only toyed with in the past. I think I can make it. I will have hundreds of shaggy primates to help keep me sane. If they fail, great evacuation insurance will fly me out to the nearest mental institution. ■

Aliza le Roux is a postdoctoral fellow in animal behaviour at the University of Michigan.

# Project: Verbivore

It's a write off.

**James Lovegrove**

UK SECRET — CLEARANCE LEVEL  
■■■■ EYES ONLY  
[NEWLY DECLASSIFIED]

Colonel ■■■■  
Chilton Mead Research Facility  
nr High Leversham  
Wilts  
17th March 1977

To: Sir ■■■■  
■■■■ ■■■■  
Ministry of Defence Main Building  
Whitehall  
London SW1

My Dear ■■■■,

I hope this finds you well. Life at Chilton Mead continues in its usual way, one long round of meetings, meetings and more meetings, with the occasional top-to-toe budgetary review to liven things up. June can't come soon enough, as far as I'm concerned. I only get a fortnight off but I need the break. I'm very much looking forward to attending your Silver Jubilee bash at your place in ■■■■. It's been so long since I last saw ■■■■, and of course young ■■■■, not to mention the delightful ■■■■. I hear her flute playing is coming along a treat and the music scholarship to Roedean was greatly deserved. Together, we shall all raise a bumper and toast Her Maj, God bless her. Fine woman, as I'm sure you know. Long may she reign.

Anyway, down to business. You asked me to furnish you with a report on the progress of Project: Verbivore. Sad to relate, things haven't gone so well. That isn't to say that the project hasn't been a success. It's simply that the results have proved unfeasible in practical terms.

Let me explain. As part of our continuing efforts at Chilton Mead to develop new and subtle methods of prosecuting war against our enemies, we have been looking into ways of undermining their intelligence-gathering and record-keeping capabilities. Destroy the informational infrastructure, destroy the foe — that's our somewhat unwieldy motto.

Now, the science johnnies here are, as you know, some of the brightest and weirdest boffins on the planet, and they don't come much brighter or weirder than Professor ■■■■. Can't stand the fellow, personally. One of those drawling longhairs who seems to



JACEY

resent having to work for the military to earn a crust and who has no respect for my authority — or anyone else's for that matter. The number of times I have had to reprimand him for his slovenly manner and his refusal to address me by my rank. And all I get in reply is 'Hey, cool it, man' or some other such ghastly, slack-jawed modernism.

However, his Verbivore concept, which you were understandably very taken with when I briefed you on it last year, seemed to be the perfect solution to the matter at hand: a bacillus that eats words.

Don't ask me to explain how it works. I honestly have no idea. Professor ■■■■ spoke in terms of engineering microbes with a taste for one particular foodstuff, much as certain lichens will grow only on certain trees and certain moulds only on certain cheeses. Most of this stuff is right over my ■■■■, more akin to magic than science. I just accept it now, after 11 years in charge at a place where white-coated whizz-kids can make graffiti sing and fleas dance the hula.

The Verbivore bacillus consumes the written word with locust-like voraciousness, leaving nothing but a black stain behind. 'Verbivore spoor,' Professor ■■■■ called it, with one of those sloppy, lopsided grins of his. (I refused to grace him with a laugh but I thought it quite a witty turn of phrase nonetheless.) Once released on a target piece of text, the bacillus grows and multiplies at speed, munching through words at the rate of two per hour on average, meaning it can eradicate the best part of a paragraph in a day.

The only trouble is — here's the catch — Professor ■■■■ bred it too specialized. Try as he might, he couldn't get the Verbivore to vary its diet. We tested it both here and in the field. Our best men in Moscow and Peking, agents ■■■■ and ■■■■, applied it to certain documents that passed under their noses.

They also applied it to microdot film and to blueprints. Nothing doing. Verbivore turned its nose up and starved to death.

Verbivore, you see, eats only English. It cannot stomach Cyrillic or Chinese script. It is too patriotic a bacillus, too partial to our vocabulary. It won't touch any other language, not even ones that use the same alphabet. French? Show it a DGSE memorandum and it loses its appetite in a flash. German? Instant indigestion. Italian? Not a hope.

As you can imagine, it ■■■■ therefore of no appreciable use to us. Indeed, all said and done Verbivore is pretty ■■■■ an own goal. It will erase our own documents, and those of ■■■■ American and Canadian allies, without a moment's hesitation, leaving them a patchwork of black blanks. As such it may be considered a ■■■■ and even a hazard to ■■■■ security. Consequently I have had no ■■■■ but to order the project ■■■■ and see to it that Prof ■■■■ destroys all existing samples of the ■■■■. As you can imagine he was rather miffed about this and threatened to ■■■■ his own back on ■■■■. How he'll manage that is hard ■■■■ say. I'd like to ■■■■ him try!

In any case, I'm sorry, ■■■■, that this is how things have ■■■■ out. I know you had high ■■■■ for the ■■■■. But you can see, surely, the risks of ■■■■ something like ■■■■ loose. We would render our own ■■■■ to ■■■■ more or less ■■■■! Talk about ■■■■ in the ■■■■.

Still, live and ■■■■, eh? Back to the old ■■■■.

Yours ■■■■,  
■■■■ ■■■■

James Lovegrove is the author of more than 20 books including *Days*, *Untied Kingdom* and *Provender Gleed*, and writes regular book reviews for the *Financial Times*.